



Proceedings

الندوة الدولية للوالب وتكنولوجيا المعلومات

# International Conference on Web and Information Technologies

Sidi Bel Abbès, Algeria

29-30 June, 2008

<http://www-inf.univ-sba.dz/icwit>



Sidi Bel Abbès University



Evolutionary Engineering And  
Distributed Information Systems



National Agency for University  
Research Development

*Proceedings Of  
The 1<sup>st</sup> International Conference on Web and  
Information Technologies*

*June, 29-30, 2008  
Sidi Bel Abbès , Algeria*

*Sponsored By*



*Evolutionary Engineering and  
Distributed Information System  
Laboratory*



*Djillali Liabes University  
of Sidi Bel Abbès*



*National Agency of University  
Research Development*

# ICWIT'o8

*International Conference on Web*

---

*and Information Technologies*



## *ICWIT 2008 General Chair's Welcome Message*



**I**t is our pleasure to present the proceedings of the first International Conference on Web and Information Technologies (ICWIT'2008), held in Sidi Bel-Abbes, Algeria, June 29-30 2008.

ICWIT Conference wants to be a convivial space and privileges discussions meetings and exchanges of results, between Algerian researchers and foreigners who aim to gather researchers around debates on recent topics of Web and Information Technologies. Another conference objective is to able young researchers of the field to present their work at the whole of the community. We would like to maintain these emphases for ICWIT, in this year and in future conferences, as awareness of progress in related fields is important for all specialists in Web and Information Technologies.

We had an overwhelming number of paper submissions for this first edition. This was attributed to the very strong technical committee that worked so hard to make the conference advertisement reach so many people. Also the strong commitment of many of track chairs was outstanding. After the submission deadline, they worked so hard with their TPC members to turn around a very high standard review that led to the strong technical program before you. We would also like to express our appreciation to all contributors for their cooperation and efforts, to all reviewers for their time and valuable comments, and to all members of the conference committees for giving generously of their time.

We are very proud of the technical program we have and assure you that you will find it very stimulating, informative and inspiring. In addition to the great technical sessions, we have 4 high-profile keynote speakers. I invite you to attend and interact with the keynote speakers as they are all distinguished in their fields.

We thank our sponsors – Djillali Liabes University, Evolutionary Engineering and Distributed Information Systems Laboratory and National agency of university research development – for making the conference and these proceedings available.

We invite all of you to join us and interact with experts in the web and information technologies field from all around the world. This will be a memorable experience you will never forget.

**Mimoun MALKI , General Chair, ICWIT 2008**

**Djillali Liabes University, Sidi Bel-Abbes, ALGERIA**

**June, 2008**



## ICWIT 2008 Program Chair's Foreword



**W**e are pleased to greet the First International Conference on Web and Information Technologies 2008, ICWIT'08, organized by the **EEDIS laboratory of the Department of Computer Science, Djilali Liabès University** (Sidi-Bel-Abbès, Algeria).

The different papers that have been carefully selected are dedicated to the fast growing field of Information Systems Design and last generation Web Technologies and Services. Having in mind the extreme importance of these research areas, the **EEDIS** laboratory has promoted this conference as a forum for scientists and researchers, both in the university and in the industry, to discuss and exchange ideas, experiences, research results and applications about all aspects of Information and Web and Information Technologies and Services, and to promote promising cooperation between the industry and the university.

We are pleased to notice that 154 submissions have been received for a number of foreign countries, including Morocco, Tunisia, France and Georgia, covering different areas of Web and Information Technologies. Referees belonging to the conference technical committee carefully reviewed all papers, and only 51 papers have been accepted for oral presentation, and 13 papers for poster.

Here down are the main topics as well as statistics on the number of accepted papers:

Topics	Number of accepted papers
Web Services	06
Automatic processing of natural language	06
E-Learning	08
Software Engineering	12
Ontologies engineering & reverse engineering	05
Mobile & Wireless technology	06
Geographic information Systems	03
Intelligent System	05

We hope that the conference satisfy your expectations and hope you enjoy time with us in Sidi Bel Abbes, Algeria. The committees worked hard for not just days but months to reach the quality level that do satisfy our guests, and hoping we did succeed in that.

Finally, we hope that this conference will be the first of a long and lasting series.

**Prof. M.K. Rahmouni , Program Chair of ICWIT 2008**

**Es-Sénia University, Oran, ALGERIA**

**June, 2008**

## *Honourable Chair of The conference*

Prof., Dr. Abdennacer TOU, Hear of Djillali Liabès University

## *Co- Honourable Chair of The conference*

Dr. Samr HADJERI, Hear of Engineering Faculty

## *Steering Committee*

Dr. Mimoun MALKI

Dr. Ahmed LEHIRECHE

Dr. Sidi Mohamed BENSLIMANE

Dr. Ahmed BOUACHRIA

## *Scientific Committee*

### **President**

Pr. Mustapha Kamel RAHMOUNI, Université Oran

### **Members:**

AMGHAR, Youssef	Lyon, France
BALLA, Amar	Alger, Algérie
BELDJILLALI, Bouziane	Oran, Algérie
BELDJILLALI, Tarik	Moncton, Canada
BELLAHSENE, Zohra	Montpellier, France
BELLATRECHE, Ladjel	Poitiers, France
BENATALLAH, Boualem	Sydney, Australie
BENHARKAT, Nabila	Lyon, France
BENMOHAMED, Med	Constantine, Algérie
BENSLIMANE, Djamel	Lyon, France
BENSLIMANE , Sidi Med	Sidi Bel Abbes, Algérie
BENYETTOU, Aek	Oran, Algérie
BETTAZ Mohamed	Alger, Algérie
BHARGAVA, Bharat	Purdue, USA
BHOWMICK, S. Saha	Nanyang, Singapour
BOUACHRIA, Ahmed	Sidi Bel Abbes, Algérie
BOUFAIDA, Mahmoud	Constantine, Algérie
BOUSSAID, Omar	Lyon, France
CHIKH, Azzeddine	Tlemcen, Algérie
CHIKH, Mohamed Amine	Tlemcen, Algérie
CHIKHI, Salim	Constantine, Algérie
CHOUARFIA, Abdellah	Oran, Algérie

DARMONT, Jérôme	Lyon, France
DJOUADI, Mahieddine	Poitiers, France
DRIAS, Habiba	Alger, Algérie
FEKI, Kamel	Sfax, Tunisie
FERAOUN, Med Kamel	Sidi Bel Abbes, Algérie
FLORY, André	Lyon, France
GAFFOUR, Abdelkader	Sidi Bel Abbes, Algérie
GARGOURI, Faiez	Sfax, Tunisie
GIACOMETTI, Arnaud	Tours, France
HAFFAF, Hafid	Oran, Algérie
KHOLLADI M. Khireddine	Constantine, Algérie
KOUNINEF, Kacem	Oran, Algérie
LAKRIB, Mustapha	Sidi Bel-Abbes, Algérie
LASKRI, Mohamed Tayeb	Annaba, Algérie
LEHIRECHE, Ahmed	Sidi Bel Abbes, Algérie
MAAMAR, Zakaria	Dubai, EAU
MADRIA, Sanjay	Missouri-Rolla, USA
MALKI, Mimoun	Sidi Bel Abbes, Algérie
MARCEL, Patrick	Tours, France
MESSABIH, Belhadri	Oran, Algérie
MISSAOUI, Rokia	Montréal, Canada
MOKADDEM, Abdelillah	Compaq, USA
MONDAL, Anirban	Tokyo, Japon
REYNAUD, Chantal	Paris, France
SCHNEIDER, Michel	Aubière, France
SELLAMI, Mohamed	Annaba, Algérie
SELOUANI, Sid-Ahmed	Moncton, Canada
SIMONET, Michel	Grenoble, France
SLIMANI, Yahia	Tunis, Tunisie
THO, Nguyen Manh	Vienne, Autriche
WIDDOWS, Dominic	Google ,USA
YOUSFATE, Abderahmane	Sidi Bel-Abbes, Algérie
ZEGOUR, Djamel Eddine	Alger, Algérie
ZEMIRLI, Zouhir	Alger, Algérie
ZIADI, Djelloul	Rouen, France

## *Organisation Committee*

### **President**

Dr. Mimoun MALKI

### **Secretariat**

Dr. Sidi Mohamed BENSLIMANE

Mr. Sofiane BOUKLI HACENE

### **Membres**

Mr. Ahmed MENAD, Secrétaire Général de l'Université de Sidi Bel abbés

Dr. Ahmed LEHIRECHE

Dr. Sidi Mohamed BENSLIMANE

Dr. Ahmed BOUACHRIA

Dr. AEK GAFFOUR

Dr. Mustapha LAKRIB

Dr. Mohamed Kamel FERAOUN

Dr. Omar BOUABDALLAH

Dr. Zakaria ELBERRICHI

Dr. Abderrahmen YOUSFATE

Mr. Mohamed Ismail ARRAR

Mr. Mustapha TOUNSI

Mr. Sofiane BOUKLI HACENE

Mr. Adil TOUMOUH

Mr. Mohamed TAIEB BRAHIM

Mr. Abdelmalek AMINE

Mr. Salim CHIALI

Mr. Mohamed BENTALLAH

Mme. Fatima Zohra HADJAM

Mr. Djamel AMAR BENSEBER

Mr. Mohamed MOKHTARI

Mr. Mohamed BENHAMOUDA

Mr. Mourad ZARHOUNI

Mr. Mourad KLOUCHE

Mlle. Houda ABBAD

Mlle. Afaf MERAZI

Mme. Cherifa DAHMANI



## *Table Of Contents*

# Table of contents

Message from the General Chair	I
Message from the Program Chair	II
Steering Committee	III
Scientific Committee	III
Organization Committee	V
<b>Keynote 1</b> Innovations technologiques au sein du projet de l'université virtuelle algérienne	XII
<b>Keynote 2</b> From Designing Data Warehousing Applications to Tuning	XIII
<b>Keynote 3</b> Les ontologies dans le domaine de la médecine et la santé	XIV
<b>Keynote 4</b> Mobile Multimedia Handsets based on STM Nomadik® Application Processors	XV

## Session 1: E-learning I

Le mobile learning pour la compréhension orale des langues Etrangères <i>Mahieddine DJOUDI</i>	01
Recommandation automatique des objets d'apprentissage aux apprenants en utilisant SVM <i>El Amine OURAIBA, Azeddine CHIKH, Abdelmalik TALEB-AHMED</i>	07
Architecture distribuée d'une plate forme du E-learning basée sur les Web services <i>Hakim AMROUCHE, Amar BALLA, Djamel CHARCHAL, Adel BOUKHEDR</i>	14
Architecture d'un entrepôt pédagogique et sa modélisation sémantique <i>Nawel ILES, Azeddine CHIKH, sidi mohammed CHOUTI</i>	20

## Session 2: Web Services

Bringing semantics to web services: Model driven approach <i>Djamel AMAR BENSABER, Djamel BENSLIMANE, Mimoun MALKI</i>	27
Vers une plateforme de la modélisation agile et de l'analyse des processus métier <i>Mohamed BOUKHEBOUZE, Youssef AMGHAR, Aïcha-Nabila BENHARKAT</i>	34
Agents mobiles et web services pour le E-Commerce <i>Linda ZEGHACHE, Aouaouche EL-MAOUHAB, Sanaa MEDJEDBA, Habiba BOULEFAAT</i>	40
Heavy-tailed distribution and QoS in distributed Web server systems <i>Fouzi SEMCHEDINE, Louiza BOUALLOUCHE-MEDJKOUNE, Djamil AÏSSANI</i>	44

Application des Métriques de Similarité sémantique pour la Composition des Services Web <i>fethallah HADJILA, M.A. CHIKH, Amine BELABED, Belkhir DJABARI, maamar KHATER</i>	51
--	----

Editeur SCORM en web services conforme à la norme IMS GWS « Instructional Management Systems General Web Services » <i>Faiza MEDJEK, Aouaouche EL-MAOUHAB, Radia ADJERAD, Nahla GHEHRIA</i>	57
--	----

### *Session 3: Ontologies engineering & reverse engineering*

Une extension de l'ontologie de domaine pour supporter la multitude de contextes lors de la spécification des besoins <i>Achraf MTIBAA, Wassim JAZIRI, Faïez GARGOURI</i>	66
--	----

Classification de techniques d'appariement d'ontologies pour le Web sémantique <i>Samira OULEFKI, karima AKLI ASTOUATI</i>	72
---	----

Un Index de Contenu et de Structure pour les Documents XML <i>Fouad DAHAK, Mouhand BOUGHANEM, Amar BALLA,</i>	78
--	----

La réingénierie de l'ontologie : vers une nouvelle approche <i>Ghizlène SOULIMANE, Mimoun MALKI, Sidi Mohamed BENSLIMANE</i>	84
---	----

Appariement Ontologique <i>Aicha BOUBEKEUR, Mimoun MALKI</i>	91
---	----

### *Session 4: Software engineering I*

Méta-fédération : une architecture pour l'assemblage d'Eléments logiciels <i>Khaled SELLAMI, Mohamed AHMED-NACER</i>	98
---	----

Structural and Behavioural Composition Aspects in Component-Based Architectures <i>Abdelkrim AMIRAT, Mourad OUSSALAH</i>	103
---	-----

Une Architecture basée Agent pour la Gestion de l'Interopérabilité des Processus Métiers <i>Yasmina RAHMOUNE, Mahmoud BOUFAIDA</i>	112
---	-----

Approche distribuée pour la classification supervisée <i>Moez BEN HAJ HMIDA, Yahya SLIMANI</i>	118
---	-----

Approche Bond-Graph pour la modélisation des systèmes d'information (BGSi): nouvelle approche orientée communication <i>bouazza ABED BAHTOU, Mustapha Kamel RAHMOUNI</i>	125
---	-----

Proposition d'une Méthode pour Améliorer la Qualité d'une Conception Orientée Objet <i>Lilia GRATI, Mohamed TMAR, Faïez GARGOURI</i>	132
---	-----

### Session 5: Geographic information Systems

Vers un système d'unification des systèmes d'informations géographiques <i>Fethi ABDICHE, Baghdad ATMANI</i>	140
Utilisation d'un SIG pour l'Évaluation des caractéristiques physiques d'un bassin versant et leurs influences sur l'Écoulement des eaux (bassin versant d'Oued El Maleh, Nord-Ouest d'Algérie) <i>Nadjla BENTEKHICI, Zine eddine ZEBBAR</i>	147
A Tool for Semantic Integration of Geographic Ontologies <i>khelifa DJERRIRI, Mimoun MALKI</i>	154

### Session 6: Natural language processing

Classification des textes arabes basée sur l'algorithme des arbres de décision <i>Fouzi HARRAG, Aboubekour HAMDICHERIF</i>	164
Modèle p-contexte de classe pour la génération automatique des phrases arabe <i>Si Lhousine AOURAGH, Ailal JELLOUL, Abdellah YOUSFI</i>	170
English-French Cross-Language Text Categorization <i>Mohamed Amine BENTAALLAH, Mimoun MALKI</i>	175
Lexicon KalimNet : Calcul automatique de la synonymie À partir d'un dictionnaire monolingue arabe <i>Amine AEK ZOUAOUI, Ahmed LEHIRECHE, Mimoun MALKI</i>	180
Utilisation des N-Grammes dans la Classification Automatique Non supervisée de Textes <i>Amine ABDELMALEK, Zakaria ELBERRICHI, Michel SIMONET, Mimoun MALKI</i>	186
Using Word Space Models for Enriching Multilingual Lexical Resources and Detecting the Relation Between Morphological and Semantic Composition <i>Adil TOUMOUH, Dominic WIDDOWS, Ahmed LEHIRECHE</i>	195

### Session 7: Software engineering II

Minimisation des Désagréments dans les Clusters Agrégés <i>Rouba BAROUDI, Safia NAIT BAHLOUL, Youssef AMGHAR</i>	202
Une Approche Cognitive Centrée Utilisateurs pour l'Élicitation des Besoins dans les Organisations Distribuées <i>Mohamed AMROUNE, Nacereddine ZAROUR</i>	208
Relevé et classification automatique des dégradations des routes bitumées <i>Younes GUELLOUMA, Hadda CHERROUN</i>	214
Dynamic Load Balancing Strategy for Grid Computing <i>Belabbas YAGHOUBI, Meriem MEDEBBER</i>	220
Extension conceptuelle de méta-modèles XML et UML pour une transformation bidirectionnelle de modèles <i>Ludovic MENET, Myriam LAMOLLE</i>	228
Le diagnostic des fautes multiples <i>Lotfi BOUZIANI, Hafid HAFFAF</i>	237



## Session 8: Mobile & Wireless technology

Enhancing DSR Protocol in MANETS <i>Sofiane BOUKLI HACENE, Ahmed LEHIRECHE, Ahmed MEDDAHI</i>	245
Amélioration des Interactions entre les Protocoles Transport et MAC dans les MANET <i>Sofiane HAMRIOUI, Mustapha LALAM</i>	251
Une Nouvelle Approche pour l'Economie d'Energie dans les Réseaux Sans Fil Mobiles <i>Sofiane HAMRIOUI</i>	258
Power control issues in mobile ad hoc networks <i>Kadri BENAMAR, Mohamed FEHAM</i>	265
Protocole de reprise efficace pour applications ubiquitaires dans les réseaux mobiles à infrastructure <i>Nabila LABRAOUI, Makhlouf ALIOUAT</i>	272
Node disjoint multipath routing protocol supporting quality of service in wireless sensor networks <i>Zahia BIDAÏ, Bouabdellah KECHAR, Hafid HAFFAF</i>	281

## Session 9: E-learning II

Modélisation de l'apprenant dans les systèmes hypermédias adaptatifs <i>Amel BEHAZ, Mahieddine DJOUDI</i>	288
Vers une adaptabilité du scénario d'apprentissage à base d'indicateur de parcours <i>Samia Ait-Adda, Nabila BOUSBIA, Amar BALLA</i>	294
Concevoir un tuteur intelligent collectif et adaptatif <i>mohamed HAFIDI, taïer BENSBAÏ</i>	300
Processus d'adaptation des activités et des documents dans un cours en ligne <i>Mokhtaria BEKKAOUI, Azzedine CHIKH</i>	309

## Session 10: Intelligent Systems

Prédiction de la structure des protéines par l'approche neuronale <i>Hafida BOUZIANE, Belhadri MESSABIH, Abdellah CHOUARFIA, Ghania KHENSOU</i>	316
Vers un système d'aide à la décision basé sur les connaissances cruciales <i>Mohamed TURKI, Faïez GARGOURI</i>	323
Vers la conception d'un modèle de classification par automates d'arbres <i>Souad Taleb ZOUGGAR, Baghdad ATMANI</i>	329
Vérification et Reconnaissance des signatures manuscrite hors-ligne en utilisant les SVMs Multi-classes et les Réseaux De Neurones A Impulsion <i>Souad BELHIA, Mohamed BENYETTOU, Ahmed LEHIRECHE, Réda ADJOUDJ</i>	336
Fuzzy Set Membership Function Construction Probabilistic Model for Applications in WWW <i>Tamaz GACHECHILADZE, Tatiana KISELIOVA, Magda TSINTSADZE</i>	343

## Posters

Towards new generation of information search tools on the web using ontologies <i>abderaouf Kouhouf</i>	349
Proposition d'un modèle pour ordonnancement d'un système automatisé de production par intensification de chromosomes <i>Djamila BOUHALOUAN, Nassima AISSANI, Bouziane BELDJILALI</i>	350
Modélisation d'un système d'information assurant l'interopérabilité des appareils mobile <i>Derdour MAKHLOUF, Nacira GHOUALMI-ZINE, Philippe ROOSE</i>	351
Le Multicast dans un Réseau Cellulaire <i>Nassima ADJISSI</i>	352
Informatique mobile: localisation d'utilisateurs <i>Zohra SLAMA</i>	353
Electronic corpora and computational linguistics <i>Mohamed GRAZIB, Bel Abbés NEDDAR</i>	354
Un Méta-Modèle Ontologique pour la Représentation de Domaines d'Enseignement en E-learning <i>farida BOUARAB-DAHMANI, Malik SI-MOHAMMED</i>	355
Opération de visioconférence en Algérie: Vers quelle pratique pédagogique? Classique ou socioconstructiviste ? <i>Hadda CHERROUN, Mouloud BOUZOUAD</i>	356
An Agent-based Architecture for an Intelligent Decision Support: Application to distributed real time Scheduling <i>Noria TAGHEZOUT, Karim BOUAMRANE</i>	357
Learning Domain Ontology for Semantic Web Services <i>Djamila BERRAMDANE, Djamel Amar BENSABER, Mimoun MALKI</i>	358
Optimisation de la consommation d'énergie avec control de topologie dans les réseaux de capteurs sans fil <i>K FELLAH, B KECHAR</i>	359
Semantic Web Services: A reverse engineering approach for the WSMO Specifications <i>Houda EL BOUHISSI, Mimoun MALKI, Djelloul BOUCHIHA</i>	360
Des Mécanismes de Tolérance aux Fautes pour un Réseau de Capteurs Sans Fil <i>Khelifa BENAHMED, Madjid MERABTI, H HAFFAF, Larbi ABDELMADJID</i>	361

## Keynotes

---



**Mahieddine Djoudi**

Equipe IRMA & Laboratoire SIC – XLIM, UMR CNRS 6172  
Université de Poitiers, Bât SP2MI, Téléport 2,  
Bvd Marie et Pierre Curie  
BP 30179 86962 Futuroscope Chasseneuil Cedex, France  
[mahieddinedjoudi@univ-poitiers.fr](mailto:mahieddinedjoudi@univ-poitiers.fr)



**Résumé :** L'apprentissage médiatisé est devenu une réalité. Il existe dans le monde un grand nombre de plates-formes et d'environnements pour l'apprentissage sur le Web où de professeurs sont impliqués dans la médiation pédagogique par les technologies de l'information et de la communication (TIC). Ces technologies constituent des ressources appréciables pour réaliser des applications à caractère pédagogique et permettent de mettre à la disposition des enseignants des outils pour répondre aux exigences des apprenants et à leurs attentes. La complexité grandissante de l'environnement technique et les difficultés d'adaptation de l'apprenant, imposent une nouvelle vision sur la manière de concevoir les environnements d'apprentissage et d'assurer un meilleur soutien à l'apprenant.

C'est dans ce contexte que se situe notre travail de recherche qui a pour objectif de définir des stratégies innovantes pour favoriser l'apprentissage par les TIC et d'aider au mieux l'apprenant mais aussi l'enseignant dans l'usage de ces environnements. A travers le projet de l'université virtuelle algérienne AVUNET, nous avons axé nos activités de recherche dans les directions suivantes :

- Apprentissage des langues étrangères (compréhension orale et écrite) ;
- Assistance intelligente de l'apprenant par les systèmes tuteurs intelligents ;
- Modélisation de l'apprenant en utilisant plus particulièrement les technologies des agents intelligents et du Web usage mining ;
- Conception d'environnement de type laboratoire virtuel en sciences expérimentales ;
- Indexation du contenu multimédia pour l'apprentissage en exploitant le langage SMIL et les ontologies ;
- Conception des systèmes hypermédias adaptatifs ;
- Prise en compte des contraintes de mobilité et d'accessibilité en apprentissage médiatisé ;
- Description et annotation sémantiques à base d'ontologies des ressources pédagogiques sur le Web ;
- Expansion de requêtes par thésaurus pour la recherche de contenu pédagogique sur le Web ;
- Mise en place et expérimentation d'une plate-forme pour l'enseignement de la maîtrise de l'information ;
- Développement de stratégies innovantes pour l'évaluation d'usage des environnements d'apprentissage

**Mots clés :** Analyse des logs, Baladeur numérique, Maîtrise de l'information, Ontologie, Systèmes auteurs, Thésaurus, Web usage mining



## *Les ontologies dans le domaine de la médecine et la santé*

**Michel Simonet**

TIMC - IN3S (Institut de l'Information et de l'Ingenierie de Sante),  
Faculte de Medecine  
38700 LA TRONCHE , France  
Michel.Simonet@imag.fr

**Résumé :** Les ontologies sont devenues le paradigme contemporain de la représentation des connaissances et aujourd'hui il n'est pas un domaine où elles n'aient une place. Depuis les catégories d'Aristote, organisées en arbre par Porphyre au 3ème siècle, le besoin d'organiser la réalité sous la forme de classes nommées et reliées entre elles par des liens sémantiques est devenu de plus en plus présent. L'introduction des taxonomies a été un facteur important de la révolution scientifique qui a permis l'émergence de la science moderne. Aujourd'hui, l'usage du terme « ontologie » est sans doute devenu abusif, mais il recouvre un besoin réel qu'il est nécessaire de mieux qualifier. En attendant une hypothétique définition consensuelle des ontologies, ou, mieux, une ontologie des ontologies qui organise leurs différentes acceptions et leurs différents usages, cet exposé fera le point de la situation actuelle des ontologies dans le domaine de la médecine et de la santé.

Les aspects historiques sont importants et la première classification médicale (classification des causes de décès, par J Bertillon), à la fin du 19ème siècle, est déjà révélatrice d'une question qui reste encore posée aujourd'hui : une ontologie est-elle construite en vue d'un usage particulier ou a-t-elle vocation d'universalité ? La réponse varie selon le milieu où elle est posée, celui de la philosophie ou celui de l'ingénierie (ou, plus largement, celui des praticiens). Nous pensons que les deux réponses ne sont pas exclusives l'une de l'autre et que différentes formes ou catégories d'ontologies ont leur place et leur intérêt, mais qu'il est important de bien les caractériser pour savoir de quoi on parle quand on emploie le terme « ontologie » dans un contexte particulier. Nous n'apporterons pas de réponse à cette question mais nous poserons certains termes du problème, en particulier à travers l'examen des liens qui sont faits entre ontologie et terminologie, et entre ontologie et schéma de classes au sens UML, par exemple.

Nous examinerons la question de la conception d'une ontologie et quelques domaines où les ontologies prennent une place importante : les bases de connaissances, la recherche d'information, l'intégration de données hétérogènes, l'extraction des connaissances. Nous terminerons en faisant un tour des réalisations auxquelles est aujourd'hui attaché le mot « ontologie » dans le domaine de la médecine et de la santé : classifications médicales, thesaurus, UMLS, et enfin l'initiative OBO (Open Biology Ontologies), qui représente un courant majeur dans ce domaine.

**Ladjet Bellatreche**

LISI/ENSMA – Poitiers University  
Futurocope 86960 France  
bellatreche@ensma.fr



**Abstract:** A data warehouse (DW) integrates huge amount of data of detailed and current data across entire organizations and enables different forms of decision making from the same database. DWs tend to be extremely large, for instance, the DW of General Motors Corporation modelled using a star schema exceeds 1.5 terabytes, and its fact table has more than 2 billions of instances. With terabytes and petabytes of data in the warehouse, complex queries can slow down performance for all decision makers and the task of managing this warehouse becomes difficult. A DW is usually modelled with a relational schema (star schema, snow flake schema). A star schema consists of a single fact table that is related to multiple dimension tables via foreign key joins. Dimension tables are relatively small compared to the fact table and rarely updated. They are typically denormalized so as to minimize the number of joins required to evaluate OLAP queries. Note that each dimension table contains several attributes usually used by OLAP queries. Queries defined on star schemas called, star join queries, are characterized by: (i) a multi-table join among a large fact table and dimension tables, (ii) each of the dimension tables involved in the join operation has multiple selection predicates and (iii) no joins between dimension tables. These queries are time consuming, since they are performed using complex operations like joins, selections and aggregations. Joins are well known to be expensive operations, especially, when the involved relations are substantially larger than the main memory.

Optimization and tuning in data warehouses are the processes of selecting adequate optimization techniques in order to make queries and updates run faster and to maintain their performance by maximizing the use of data warehouse system resources. To continuously improve query performance, two main phases are required: physical design and tuning. In the first phase, data warehouse administrator selects optimization techniques such as materialized views, advanced index schemes, denormalization, vertical partitioning, horizontal partitioning and parallel processing, etc. Generally, this selection is based on most frequently asked queries and typical updates. Physical design generates a configuration C with a number of optimization techniques. This configuration should evolve, since data warehouse dynamically changes during its lifetime. These changes necessitate a tuning phase so as to keep the performance of the warehouse from degrading. Changes may be related to the content of tables, sizes of optimization structures selected during physical design (materialized views, indexes and partitions), frequencies of queries/updates, addition/deletion of queries/updates, etc. The role of the tuning phase is to monitor and to diagnose the use of configuration C and different resources assigned to C (like buffer, storage space, etc). For instance, if an optimization technique, like an index, is not used by the whole workload, it will be dropped by a tuning tool and might be replaced with another technique.

In this talk we review in details most used optimisation techniques supported by commercial database systems. For each technique, its selection problem and some selection algorithms are described. Secondly, we present a tool developed at LISI for assisting data warehouse administrator during her/his administration task. Finally, some open issues regarding physical design and tuning are presented.

## *Mobile Multimedia Handsets based on STM Nomadik® Application Processors*

**Abdelillah Mokkedem**

STMicroelectronics,  
Wireless Multimedia Division, Rabat Morocco.  
abdelillah.mokkedem@st.com

**Abstract:** The mobile world is evolving into a personal experience with a wide range of multimedia capabilities, including capturing, playing, sending and receiving multimedia content, playing 3D games, finding a location, and even more tomorrow. ST's Nomadik SoCs (system-on-chip) and platforms are a culmination of multimedia and power-saving breakthroughs to continue this trend towards a greater end-user experience. STMicroelectronics (ST) has been working for the last several years in partnership with other leaders including Nokia, ARM and AMD in order to extend its portfolio in low-power multimedia solutions for next-generation mobile products. Highly integrated devices, complemented with customizable reference designs, high-performance and low power ARM processors and full software packages, enables handset manufacturers to develop optimized solutions with the best combination of price, performance, and power consumption, in the shortest possible time.

The second generation of Nomadik® family, STn8815, in CMOS 90nm production, provides ultra-low power consumption, open platform strategy, exceptional video and audio quality, very small footprint and scalable performance for cutting-edge handheld consumer products. The STn8815 implements a multiple-core distributed architecture together with a series of breakthroughs in video-coding efficiency and inventive algorithms, enabling smart phones, feature-rich portable multimedia devices and players, Internet tablets, portable navigation devices and mobile TVs to play broadcasted media content, record images and video clips, and perform bidirectional audio-visual communication with other systems in real time. STn8815 is already available to leading handset makers to develop the next-generation of mobile multimedia products with a choice of leading software platforms, including Symbian OS/S60, Linux, Windows Mobile and Windows Embedded CE. The STn8815 now also offers a wider range of packaging options, plus a boost in clock speed, up to 393MHz, of its ARM9 embedded microprocessor, equipped with a level 2 cached memory.

The third generation of Nomadik® family STn8820, in CMOS 65nm production, provides even greater performance with more features. It's based on the ARM1176-JZFS embedded microprocessor, a multi-level high-speed interconnect, a 3D graphical accelerator and embedded security hardware.

The next Nomadik mobile chip, which is under development today, will be produced in CMOS 45nm, will use a dual-core ARM Cortex running at 600MH, a network-on-chip (NOC) interconnect running at 200MH, and several multimedia DSPs running at multi-clock and multi-power domains, combined with a 3G GSM modem all integrated in a single-chip (SoC).

The presentation will focus on the architecture of the Nomadik SoCs, power-management units, mobile multimedia cores, and connectivity peripherals. With CMOS devices shrinking in size (cmos45), dynamic power leakage becomes significant and dynamic power consumption becomes a "must to do" in the physical implementation flow. I will present the overall design flow and methodology in ST Wireless Multimedia Division and show how these low-power chips are being designed today using cutting-edge EDA tools from Synopsys, Cadence, Mentor-Graphics and others.

*E-learning I*

---





## **Le mobile learning pour la compréhension orale des langues étrangères**

Mahieddine Djoudi

Laboratoire SIC - XLIM et Equipe IRMA  
UFR Sciences SP2MI, Université de Poitiers  
Téléport 2, Boulevard Marie et Pierre Curie,  
BP 30179, 86962 Futuroscope Chasseneuil Cedex, France  
mahieddine.djoudi@univ-poitiers.fr

**Résumé.** L'objectif principal de notre travail est la conception et la mise en œuvre d'un dispositif innovant dédié à la compréhension orale de langues étrangères. Ce dispositif est basé sur la combinaison d'un véhicule d'information (le web), d'un outil technique (le baladeur numérique) et d'une méthode pédagogique adaptée. Cette démarche qui se situe dans le domaine de l'apprentissage mobile et qui vise à doter les apprenants de baladeurs numériques. Cet objet qui, dans son utilisation première sert, à écouter de la musique au format mp3, est détourné de cet usage pour en faire un outil au service de l'apprentissage des langues. Cette démarche s'appuie sur une technologie facilement utilisable et permet aux apprenants de travailler, à leur rythme, la compréhension orale d'une langue étrangère. Il s'agit de motiver les apprenants et de favoriser " l'individualisation " de l'écoute de fichiers sonores. Ces fichiers, créés par l'enseignant, sont déposés par celui-ci sur le serveur web de la plate-forme d'enseignement, facilitant ainsi aux apprenants l'accès à ces ressources disponibles en permanence en téléchargement.

**Mots-clés :** Apprentissage mobile, capacité en langues, compréhension orale, baladeur numérique, fichier MP3.

### **1 Introduction à l'apprentissage mobile**

L'apprentissage mobile ou m-learning peut être défini comme l'utilisation des supports mobiles à des buts pédagogiques. Les outils de la mobilité voyagent souvent avec les personnes [1] et parmi ces outils on peut citer le téléphone [2], le PDA ou « Personal Digital Assistant » [3], le Pocket PC [4], les tablettes PC [5], les baladeurs numériques [6], etc. Les technologies mobiles sont en train de transformer le monde éducatif. La question est de savoir comment ces technologies affectent l'environnement d'apprentissage, la pédagogie et l'éducation permanente [7]. Pour Bryan Alexander, les technologies mobiles et leur adoption par les jeunes générations sont appelées à transformer l'éducation elle-même [8].

### **2 Capacités en langue et compréhension orale**

L'utilisation d'une langue met en jeu quatre compétences (four skills, en Anglais). Deux sont du domaine de la compréhension. Il s'agit de la compréhension orale et de la compréhension écrite. Les deux dernières relèvent de l'expression (ou production) orale et écrite. Une méthodologie peut donner la priorité à une ou deux de ces compétences ou bien elle peut viser l'enseignement/apprentissage de ces quatre compétences ensemble ou selon une programmation déterminée. Dans la perspective de l'apprentissage d'une langue, on peut dire que "comprendre" signifie accéder au sens fondamental du document lu ou écouté. En termes d'activités de classe, cela signifie que l'objectif premier d'une activité de compréhension est de travailler sur le contenu et non sur des mots ou une catégorie de mots. La compréhension orale correspond à la compétence la plus fréquente et peut se résumer dans la formule "entendre et conférer un sens". Chronologiquement, c'est celle qui se met en place en premier, sauf dans des situations exceptionnelles (personnes uniquement ou d'abord confrontées à l'écrit, déficients auditifs, étude d'une langue morte, étude d'une langue en partant de l'écrit pour un autodidacte).

### **3 Contexte et motivations**

#### **3.1 Constat**

L'utilisation des appareils mobiles pour l'apprentissage des langues se développe à grande vitesse ces dernières années. Ainsi, nous assistons à une profusion de projets de recherche et de développement, de méthodologies et de publications scientifiques [9], [10]. Néanmoins, l'intérêt pour la compétence de compréhension orale reste relativement faible. Notre idée part d'un constat sur la situation de l'orale en apprentissage des langues étrangères pour proposer une démarche simple et originale qui utilise le baladeur numérique en compréhension orale chez l'apprenant.

Le déficit de pratique orale chez les apprenants s'amplifie de plus en plus. Ce phénomène résulte de plusieurs facteurs qui vont de la classe surchargée, à un scepticisme de la part des apprenants quant à la nécessité de communiquer en langue étrangère.

Ces éléments combinés font que l'oral se résume souvent à quelques réponses des apprenants les plus à l'aise face aux différentes situations de pratique de la langue proposées par l'enseignant, à l'écoute et à l'étude de quelques documents sonores.

Par ailleurs, il existe très peu d'expériences actuellement menées qui s'intéressent à l'utilisation de dispositifs nomades dans l'apprentissage des langues étrangères, notamment dans le développement de la compréhension des langues étrangères. En effet bien que nous soyons entrés dans l'ère de l'hypermédia et du multimédia, la plupart des contenus pédagogiques auxquels ont accès les apprenants sont des matériels textuels. Y compris dans les contenus pédagogiques relatifs aux langues.

#### **3.2 Motivations**

Les enseignants de langues constatent qu'il est très difficile de faire travailler l'oral en langues vivantes (classes souvent surchargées, horaires réduits, prépondérance de l'écrit dans les examens, réticences à prendre la parole à un âge difficile, manque d'assurance de l'apprenant lié à un manque de confrontation à la langue, etc.).

On peut dans certains cas avoir la chance d'avoir dans son établissement un laboratoire de langues, ou bien faire travailler l'oral individuellement ou à deux sur des ordinateurs mais il reste toujours le problème de l'accès aux outils de retour à la maison. Comment travailler l'oral quand on n'a plus accès aux outils ?

Partons de ce constat, notre idée consiste à les apprenants de baladeurs numériques pour qu'ils puissent avoir accès en "autonomie guidée" aux enregistrements sur lesquels l'enseignant se propose de les faire travailler [11]. Le travail sur le support audio peut être fait en classe de façon 'traditionnelle' ou sur PC ou en laboratoire de langues et surtout poursuivi à la maison [12].

### **4 Description de la démarche d'apprentissage**

#### **4.1 La méthode pédagogique**

La méthode pédagogique appelée ACADEME (Apport et Convergences Au-delà Des Médias) a été inventée par Jean Sabiron [13]. Son principe vise à améliorer la compréhension de la langue étrangère parlée en confrontant l'apprenant à des documents audio relatifs à une même information où les documents varient selon les accents (américain, anglais, etc.) et les registres (présentation, réaction, prolongement).

Le principe général de la méthode repose sur l'idée qui en présentant une information sous différents points de vue, registre, accents, l'apprenant doit dégager les invariants, pour être capable d'extraire les éléments cruciaux pour comprendre globalement l'information et par voie de conséquence favoriser le développement de la compréhension des langues parlées.

La démarche d'apprentissage assure ainsi :

- Une facilitation de l'écoute par des énoncés conçu spécialement ainsi que par l'apport d'éléments d'information complémentaires et une progression par niveaux phonologiques et lexicaux.
- Elargissement progressif de la représentation initiales entre les différents renoncés qui aide l'apprenant à organiser et intérioriser une représentation de la situation. L'anticipation des propositions attendus lui laisse du temps et de la capacité cognitive pour faire des hypothèses et des inférences plus fragmentaires.
- Une construction du sens à partir des variétés des messages (rapidité, styles, variations phonologiques). Ainsi d'un document à l'autre, l'apprenant élargit son angle d'audition tout en rattachant les nouvelles propositions aux connaissances préalables.

## 4.2 Diffusion des documents

La diffusion des fichiers sonores sur le serveur web de la plate-forme pédagogique conçue à cet effet, permet au public autorisé un téléchargement rapide des documents sonores. La plate-forme fournit également des instructions sur le travail à faire et de corrigés d'exercices pour l'apprenant isolé. L'écoute des fichiers-son se fait ensuite à partir de l'ordinateur lui-même, et surtout par téléchargement de ces d fichiers sur un baladeur numérique pour une écoute indépendante de l'ordinateur.

L'innovation se fait donc par deux niveaux successifs de démultiplication de la diffusion : d'abord à partir du serveur vers les ordinateurs en réseau, puis à partir de chaque ordinateur destinataire vers un nombre sans limites de baladeurs approvisionnés localement (voir fig. 1) [13].

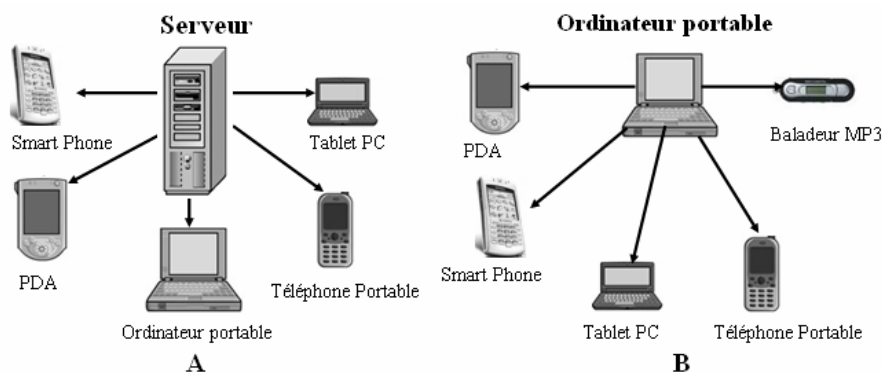


Fig. 1. Démultiplication des fichiers sons

La démarche s'appuie sur la combinaison de technologies (Informatique, Internet et Baladeur), requière peu de compétences en informatique et son coût financier est relativement modéré. Elle vise d'une part, à quantifier et qualifier les impacts liés à l'utilisation d'un dispositif innovant dédié à l'apprentissage de langues étrangères et d'autre part, à étudier les processus d'appropriation d'un outil technique spécifique faisant partie du dispositif. L'extension de l'utilisation du dispositif dans d'autres apprentissages et/ou pour d'autres types d'apprenants apparaît comme une perspective à court et moyen terme [6], [13].

## 4.3 Objectifs attendus

La démarche propose aux apprenants une situation nouvelle d'apprentissage de la langue dont un des objectifs principaux, est de mieux aider les apprenants à devenir des locuteurs capables de se faire comprendre et de progresser rapidement en langue étrangère. Elle prétend également donner une cohérence à l'apprentissage de la langue à travers une exposition démultipliée à la langue cible. Ainsi, le baladeur numérique se présente comme un outil adapté à la réalisation de ces objectifs puisqu'il permet :

- aux apprenants de se familiariser à un nouvel environnement technologique, à un nouvel espace de travail et à une méthode de travail différente intégrant les technologies de l'information et de la communication.
- de diversifier les formes d'enseignement et d'apprentissages des langues en liaison avec les réformes engagées et dans le respect des programmes nationaux.

- de proposer des situations d'apprentissages aux apprenants qui leur redonnent confiance et les motivent. En ce sens, l'utilisation du baladeur en langue contribue à une modification positive de l'attitude des apprenants dans le sens où une implication plus forte de tous est nécessaire [9].
- de développer l'autonomie des apprenants (ils ont accès aux informations et ressources de leur groupe de travail en permanence par le biais de la plate-forme) et favoriser le travail personnel régulier et constant [11].
- de modifier les habitudes de travail individuel des apprenants [14]. Des tâches spécifiques à réaliser toutes les semaines favorisent un entraînement régulier à la pratique de l'oral.

La démarche se situe résolument dans le cadre général d'une pédagogie qui vise à rendre l'apprenant le plus autonome possible et surtout, le plus actif possible : actif dans ses apprentissages, actif dans la construction de ses savoirs [15], [16]. Pour cela nous utilisons le baladeur numérique comme outil qui sert nos objectifs pédagogiques et non pas parce qu'il est techniquement un objet performant [17], [11].

## 5 Plate-forme pédagogique

La plate-forme pédagogique de notre dispositif est une application « full web » qui fournit aux trois principaux utilisateurs (enseignant, apprenant, administrateur) un dispositif qui a pour premières finalités la mise à disposition et la consultation à distance, de contenus pédagogiques pour l'enseignement des langues, l'individualisation de l'apprentissage et le télé-tutorat (voir fig. 2) [17].

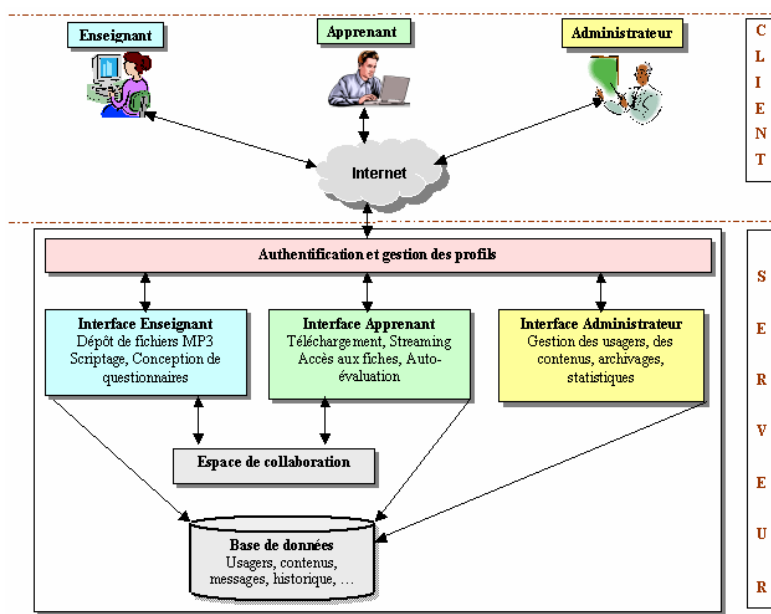


Fig..2. Architecture logicielle de la plate-forme pédagogique

Les fichiers-son sont accompagnés de documents textuels d'introduction au sujet, de contextualisation, de présentations des locuteurs étrangers, de leur variation phonologique, de façon à permettre à l'auditeur individuel de repérer les caractéristiques de la langue parlée.

### 5.1 Interface enseignant

La plate-forme pédagogique permet à l'enseignant, via une interface dédiée, de mettre à la disposition des apprenants une masse considérable de documents audio numériques compressés, d'excellente qualité d'écoute. Ces documents peuvent être créés par les enseignants ou bien récupérés sur Internet. L'interface permet aussi à l'enseignant de décrire le plus complètement possible les fichiers sonores. Les informations relatives à chaque fichier sont entre autres : le nom, la langue, la durée, le public visé, les objectifs pédagogique attendus, la période d'accessibilité, la source, copyright, etc. Les documents ainsi préparés par l'enseignant sont placés dans la base de données située sur serveur de la plate-forme.

## **5.2 Interface apprenant**

Un tel environnement de travail met, de façon certaine, l'apprenant au centre de la tâche. Celui-ci a un objet : un document oral à comprendre, qui peut contenir des obstacles. Il a aussi des outils, de type standard, pour résoudre la situation. L'apprenant se trouve donc face à une situation qui pose problème, où il va mettre en œuvre ses stratégies préférentielles, selon le modèle psycholinguistique de la « boîte noire », puisque l'on ne préjuge pas des activités qu'il va devoir déployer pour comprendre. En admettant dès lors l'importance de disposer d'outils variés et accessibles par rapport aux difficultés que peut éprouver chaque apprenant.

L'interface apprenant permet de mettre en place une mémorisation automatique des informations liées à l'activité de l'apprenant lorsqu'il exécute un scénario sur un objet pédagogique (date et durée de chaque connexion, les fichiers MP3 téléchargés ou écoutés en streaming, les exercices d'auto évaluation, etc.). Une exploitation de ces informations par l'apprenant peut le guider dans son parcours d'apprentissage.

## **5.3 Contenu et structure du document**

Les documents d'actualité récente illustrative de la thématique linguistique et culturelle du pays de la langue. La plate-forme guide l'apprenant selon la procédure suivante :

- Présentation : une même information est présentée de plusieurs façons différentes (accents, graduation de difficultés, registre de langues, etc.)
- Réaction : présence de quelques documents de réaction à cette information.
- Extension : plusieurs documents de prolongement ou approfondissement de la problématique abordée.

Chaque document sonore est d'une durée variant entre 30 secondes et 2 minutes et l'ensemble n'excède pas les 8 minutes.

Le serveur gère l'adaptation et l'optimisation du flux et du contenu, la qualité de service. L'adaptation au réseau et au terminal doit se faire en temps réel. Des réseaux de distribution de contenus mobiles se développent, sur le modèle des réseaux de distribution de contenus (en Anglais Content Delivery Networks) de l'Internet. Aux extrémités des réseaux, au plus près de l'utilisateur, des serveurs multimédias gèrent une part de la distribution et de l'adaptation au contexte utilisateur. Il est important de garder trace des usages audio et des téléchargements de matériaux : identité, date, durée, fréquence. Ces matériaux seront programmés pour être disponibles pendant des périodes limitées dans le temps [17].

## **6 Conclusion**

Nous avons présenté dans cet article une démarche originale pour la compréhension orale d'une langue étrangère en utilisant un objet détourné de sa fonction initiale. Le baladeur en tant qu'objet nomade se rapproche beaucoup de la force du livre par ses caractéristiques de portabilité, d'accessibilité, et d'autonomie.

En perspective, la démarche vise à développer chez l'apprenant l'autre compétence orale ; à savoir l'expression orale pour lui permettre ainsi à s'exprimer en langue étrangère. La maîtrise de la langue passe par la maîtrise de l'élocution. La démarche prévoit ainsi de mettre les apprenants en situation de produire des fichiers son, fruit de leur travail, par le biais de la fonction « enregistrement » du baladeur [6]

## **Références**

1. Kadyte, V. and Akademi A.. Learning Can Happen Anywhere: A Mobile System for Language Learning. Proceedings of Mlearn 2003 Conference on Learning with Mobile Devices. Central London. UK. (2003).
2. Attewell, J., and Savill-Smith, C. Young People, Mobile Phones and Learning. London: Learning and Skills Development Agency. (2003).
3. Kneebone, R. PDAs as Part of Learning Portfolio. Proceedings of Mlearn Conference on Learning with Mobile Devices. Central London. UK. 19-20 May. (2003).



4. Holme, O and Sharples, M Implementing a student learning organizer on the pocket PC platform. Proceedings of MLEARN 2002: European Workshop on Mobile and Contextual Learning. Birmingham, UK, (2002). 41-44.
5. Willis, C. L. and Miertschin L.. Technology to enable learning II: Tablet PC's as instructional tools or the pen is mightier than the 'board!. Proceedings of the 5th conference on Information technology education. Salt Lake City, UT, USA. (2004). 153 - 159.
6. Bayon-Lopez, D. Audio Nomade: un laboratoire de langues virtuel. 3e Journée des langues vivantes: l'oral : stratégies d'apprentissage et enjeux.. 24 novembre. CDDP Gironde, Bordeaux, France. (2004).
7. Mifsud L Alternative Learning Arenas - Pedagogical Challenges to Mobile Learning Technology in Education. Proceedings IEEE International Workshop on Wireless and Mobile Technologies in Education, Växjö, Sweden. (2002). 112-116.
8. Bryan A. Going Nomadic: Mobile Learning in Higher Education. Educause, vol. 39, N° 5. (2004).
9. Norbrook H. and Scott P. Motivation in Mobile Modern Foreign Language Learning Proceedings of Mlearn 2003 Conference on Learning with Mobile Devices. Central London. UK. (2003).
10. Sharples, M. Disruptive Devices: Mobile Technology for Conversational Learning. International Journal of Continuing Engineering Education and Lifelong Learning, 12, 5/6, (2003). 504-520.
11. Little, D. Learner autonomy: Why foreign languages should occupy a central role in the curriculum. In S. Green (Ed), New perspectives on teaching and learning modern languages, 24-45. Clevedon: Multilingual Matters. (2000).
12. Farmer, M. and Taylor B. A Creative Learning Environment (CLE) for Anywhere Anytime Learning. Proceedings of the European Workshop on Mobile and Contextual Learning, The University of Birmingham, England. (2002).
13. Sabiron J. Outils techniques et méthodologiques de l'apprenant nomade en langues étrangères. Computer a primavera 2003, Biblioteca regionale – Aosta, Italia. (2003).
14. Lundin, J. and Magnusson, M. Walking & Talking - Sharing Best Practice. Proceedings IEEE International Workshop on Wireless and Mobile Technologies in Education, August 29-30, Växjö, Sweden. Pp. 71 –79. (2002).
15. Mitchell, A. Developing a Prototype Microportal for M-Learning: a Social-Constructivist Approach. Proceedings of the European Workshop on Mobile and Contextual Learning, The University of Birmingham, England (2002).
16. Zurita, G. and Nussbaum M. A constructivist mobile learning environment supported by a wireless handheld network. J. Comp. Assisted Learning 20(4), (2004). 235-243.
17. Djoudi, M. Listening Comprehension of English Second Language with Mobile Devices. Journal of Mobile Multimedia (JMM), ISSN: 1550-4646, Vol. 2, No. 2, (2006). 146-166.

## Recommandation automatique des objets d'apprentissage aux apprenants en utilisant SVM

OURAIBA El Amine<sup>1</sup>, CHIKH Azeddine<sup>2</sup>, TALEB-AHMED Abdelmalik<sup>3</sup>

<sup>1,2</sup>Département d'Informatique - Faculté des sciences de l'ingénieur  
Université de Tlemcen – Algérie

<sup>3</sup>LAMIH, UMR CNRS-UVHC8530

Université de Valenciennes et du Hainaut Cambrésis - France

<sup>1,2</sup>{amine\_ouraiba, az\_chikh}@univ-tlemcen.dz <sup>3</sup>abdelmalik.taleb-ahmed@univ-valenciennes.fr

**Résumé.** En reconnaissant son importance, la personnalisation a été adoptée dans différents systèmes du e-learning : les systèmes intelligents de tutorat ; les systèmes hypermédias adaptatifs ; et les systèmes de filtrage et de recommandation d'information. Dans ce papier nous nous intéressons à cette dernière catégorie des systèmes. En effet, nous proposons une nouvelle démarche de construction d'un système de recommandation automatique qui permet aux concepteurs des cours de sélectionner automatiquement les objets d'apprentissage adaptés aux apprenants. Cela est possible grâce à un classificateur automatique qui permet de décider si un objet d'apprentissage est pertinent ou non pour un apprenant particulier. Ce classificateur est construit par la technique SVM en utilisant une base de données qui contient les informations relatives à des étudiants de l'université de Tlemcen et les descripteurs des ressources pédagogiques en informatique. Ces étudiants et ressources sont qualifiés respectivement avec IMS LIP et LOM. Nous avons implémenté notre classificateur SVM avec l'environnement libre Weka. Les résultats du test ont montré des performances très prometteuses par rapport aux autres approches existantes telles que : PERSO, EPERS, RACOFI et le Web mining par règles d'association. Les valeurs du taux de classification, de précision et du rappel de notre classificateur SVM sont très satisfaisantes.

**Mots clés :** e-learning, SVM, personnalisation, recommandation, pertinence  
profil d'apprenant, objet d'apprentissage.

### 1 Introduction

La croissance continue des informations sur le Net rend la tâche plus difficile aux internautes de trouver ce qu'ils veulent effectivement. Alors que la personnalisation consiste à fournir aux individus les contenus et les services qu'ils désirent selon leurs caractéristiques gardées dans le modèle d'utilisateur [7].

En particulier, dans le contexte de la formation ouverte et à distance (FOAD), un environnement informatique pour l'apprentissage humain (EIAH) doit adapter le contenu et les activités aux apprenants en termes de leurs préférences, connaissances, besoins d'apprentissage ou situation d'apprentissage,...etc[5]. Pour ce faire, différents types de systèmes ont été proposés, nous retenons les *systèmes intelligents de tutorat (SITs)* qui adaptent leurs explications et stratégies d'enseignements aux apprenants en termes de leurs niveaux de connaissances et leur progression d'apprentissage ; les *systèmes hypermédias adaptatifs (SHAs)* permettent de personnaliser les présentations et les navigations aux utilisateurs ; et enfin d'autres systèmes qui ont montré leurs preuves en terme de personnalisation, se sont les *systèmes de filtrage d'information* qui sont appelés aussi *systèmes de recommandation (SRs)*. L'objectif de ces derniers est de filtrer le contenu selon les informations liées aux utilisateurs.

Cependant, les SITs et SHAs souffrent d'une pauvreté et d'une insuffisance remarquables en termes d'information sur les apprenants d'une part et sur les objets d'apprentissage (LOs) d'autre part, car la sélection de ces informations est faite d'une manière artisanale. La collecte informelle de ces informations n'est pas du tout garantissante et suffisante toute seule. Alors que cela nécessite une modélisation interprétée par des experts suivant des normes et des standards spécifiques tels que : IMS LIP, IEEE PAPI, LOM,...etc. Notons que ceci a un impact majeur sur l'efficacité de la tâche de personnalisation. Comme limites remarquables, on constate également que la complexité des SHAs est due aux différents traitements de sélection et de combinaison effectués sur les données, cela implique les problèmes de désorientation et de surcharge cognitive chez l'apprenant. Ainsi, on note que les performances et les

efficacités de ces systèmes sont fixées parce qu'ils n'ont pas l'habilité d'apprentissage durant leur utilisation tels que les classificateurs automatiques.

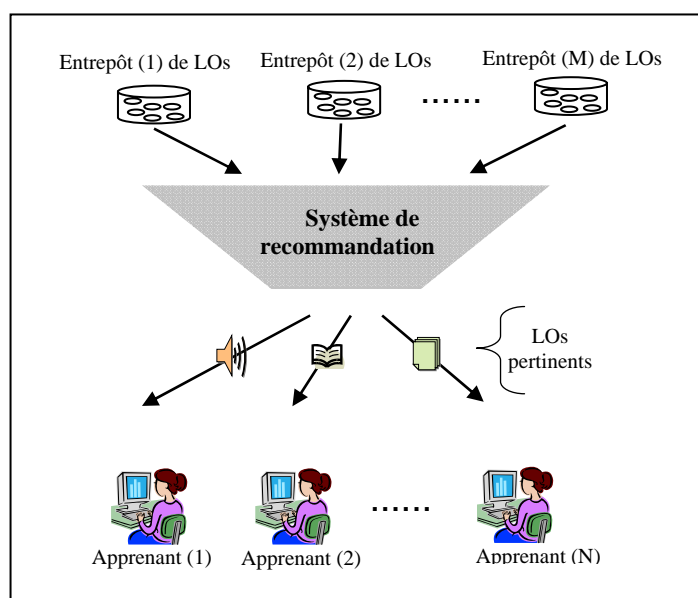
En revanche, les SRs visent à prévoir la pertinence d'un objet à un utilisateur. Ils diffèrent par des types de données décrivant les utilisateurs et les objets ou par des algorithmes utilisés pour effectuer le filtrage. D'un point de vue architectural et algorithmique, Mobasher a distingué dans [7] trois catégories des systèmes de recommandation : systèmes à base de règles ; systèmes de filtrage par contenu (cognitif) ; et systèmes de filtrage collaboratif (social).

Dans le présent article nous nous intéressons à l'utilisation des SRs dans le domaine du e-learning. Nous abordons la problématique de la recommandation automatique des LOs aux apprenants en utilisant la technique de classification « SVM ». Pour ce faire, nous nous focalisons en premier lieu sur la recommandation du contenu dans le domaine du e-learning en concentrant sur la modélisation des apprenants et des LOs. Puis, nous présentons la technique SVM. Par la suite, nous expliquons l'objectif et le principe de fonctionnement de notre système de recommandation de contenu. Dans la dernière section, nous entamons la construction de classificateur SVM en présentant les données invoquées avec leurs modèles, les phases nécessaires et enfin nous discutons les résultats du test.

## 2 Vers un Système de Recommandation des Objets d'Apprentissage aux Apprenants par SVM (SROA2)

En e-learning la population des utilisateurs est très importante et hétérogène, les utilisateurs ont des profils très divers, des rôles distincts, des besoins et centres d'intérêt différents,...etc. Pour cela, la personnalisation au sens large de terme dans un EIAH est devenu une fonction indispensable. Elle permet particulièrement de satisfaire les besoins spécifiques des apprenants, d'accomplir leurs objectifs d'apprentissage et de s'adapter aux contextes spécifiques des différentes communautés. Les SRs évitent de noyer les apprenants dans des informations qui ne sont pas pertinentes par rapport à leurs caractéristiques individuelles gardées dans le modèle d'apprenant.

Cependant, la tâche de recommandation peut être abordée comme une application du Data Mining où elle est vue comme un processus holistique et non plus comme algorithmes distincts ou types de données spécifiques. Dans ce processus, des techniques de l'intelligence artificielle trouvent leur place pour développer des systèmes de recommandation automatique. Sachant que ces systèmes font parvenir les LOs aux apprenants qui en ont besoin à partir de larges volumes d'objets générés continuellement, voire augmenter la quantité d'objets pertinents collectés à partir de différents entrepôts en ciblant les objets vraiment appropriés. La figure 1 illustre le processus de recommandation du contenu d'apprentissage aux apprenants.



**Fig.1.** Processus de recommandation des objets d'apprentissage aux apprenants.

En effet, un système de recommandation de contenu pédagogique doit contenir principalement : les profils des apprenants, les descriptions des LOs et les techniques de l'intelligence artificielle pour prédire la pertinence. De plus, les informations du contexte d'apprentissage, qui sont aussi très nécessaires, sont néanmoins généralement intégrées dans les modèles des apprenants et/ou des LOs.

## 2.1 Modélisation de l'apprenant

Le modèle d'apprenant est l'élément pivot dans les systèmes de personnalisation d'apprentissage. Pour cela, plusieurs modèles sont proposés : IMS LIP; IEEE PAPI; eduPerson; vCard; EDS; modèle de Chen et Mizoguchi; Dolog LP; FOAF;...etc. Néanmoins aucun de ces modèles ne semble avoir vraiment considéré comme la norme d'industrie, probablement parce que chacune de ces spécifications a ses inconvénients et aucune n'est généralement suffisante toute seule.

Cependant, la spécification IMS LIP récemment semble avoir gagné une acceptation plus répandue car elle est très riche et intégrale en terme d'information sur l'apprenant. IMS LIP est une collection d'informations liées à l'apprenant qui permet d'enregistrer et gérer l'historique de l'apprentissage de l'apprenant; d'engager ou non l'apprenant dans une expérience d'apprentissage; et de découvrir des opportunités d'apprentissage de l'apprenant [4]. Le modèle IMS LIP définit une structure de données en 11 catégories [4]: *Identification; Accessibility; QCL; Activity; Goal; Competency; Interest; Transcript; Affiliation; Securitykey; Relationship*.

## 2.2 Modélisation de l'objet d'apprentissage

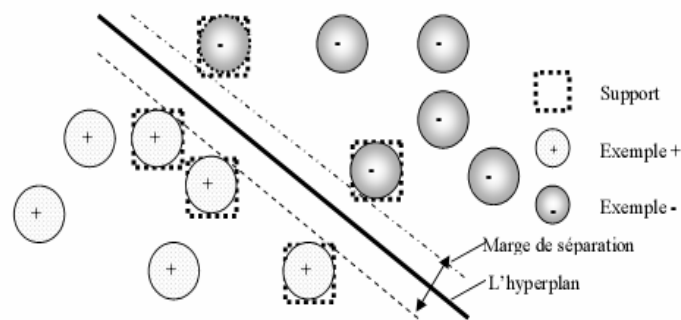
Le besoin de personnalisation dans des environnements d'apprentissage a une influence sur l'évolution d'éclaircissement des ressources pédagogiques. De nombreux standards ont été développés ces dernières années pour rationaliser la description des LOs : DCMI; ARIADNE; LOM; SCORM; EML; IMS LD;...etc. Cependant, le plus utilisé de ces standards est celui élaboré par l'IEEE en 2002 «LOM». LOM est un modèle destiné à la description de toute entité numérique ou non numérique qui peut être utilisée pour la formation [3]. Il facilite la recherche, l'évaluation, l'acquisition et l'utilisation des LOs par des concepteurs, des apprenants voire des agents logiciels. LOM se manifeste en neuf catégories [3]: *General; Life cycle; Meta-meta-data; Technical; Educational; Rights; Relation; Annotation; Classification*.

## 2.3 Une technique de classification automatique « SVM »

SVM (Support Vector Machine) [10] est une technique d'apprentissage automatique utilisée pour faire la classification [2]. Elle consiste à minimiser simultanément l'erreur empirique de classification et maximiser la marge géométrique entre les classes, «Séparateur à Vaste Marge» [8]. Parmi les avantages de SVM, son habilité d'apprentissage qui peut être indépendante de la dimension du vecteur d'entrées. La force de SVM tient à sa simplicité de mise en œuvre pour résoudre des problèmes difficiles et à ses fondements mathématiques solides. Nous retenons que les modèles SVM ont déjà montré leurs preuves dans plusieurs domaines tels que la classification du texte, d'images et la reconnaissance de locuteurs...etc. En particulier, pour notre problématique de filtrage des LOs pertinents pour les apprenants, SVM tente à séparer des cas où un LO est pertinent pour un apprenant et des cas inverses (de non pertinence) par l'hyperplan *optimal* qui garantit que l'écart entre les deux classes soit maximal en minimisant le risque d'erreur. Formellement un hyperplan a l'équation :  $\vec{w} \cdot \vec{x} - b = 0$ . Avec « $\vec{x}$ » un vecteur de position, « $\vec{w}$ » un vecteur de direction et « $b$ » le biais. On suppose :  $D = \{(x_i, y_i)\}$  le jeu d'entraînement et  $y_i \in \{\pm 1\}$  définissant la classe ou l'état positif ou négatif de l'entrée. Trouver l'hyperplan maximisant la marge séparatrice qui égale à  $(2 / \|\vec{w}\|)$  revient à résoudre le problème suivant [10] [2] [8] :

$$\text{Minimiser : } \frac{1}{2} \|\vec{w}\|^2 \quad \text{sous les contraintes : } \forall i, y_i (\vec{w} \cdot \vec{x}_i + b) - 1 \geq 0 .$$

La figure 2 montre une séparation linéaire par un hyperplan optimal.



**Fig.2.** SVM est un classificateur à base d'hyperplan.

Le cadre en pointillé en figure 2 représente un vecteur de support qui désigne tout exemple dont la distance à l'hyperplan est égale à la marge de séparation. Pour les cas non linéairement séparables, on utilise une fonction *noyau* pour traduire les données représentées dans l'espace d'origine vers un autre espace dit de « redescription » [8].

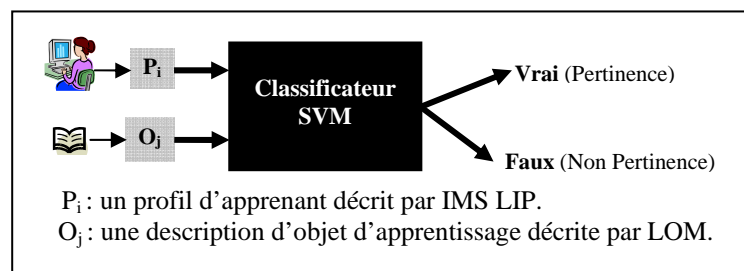
## 2.4 Objectif de SROA2

Nous nous intéressons à la réalisation d'un système de recommandation automatique des LOs aux apprenants, baptisé «SROA2». L'utilisation de ce système est destinée aux concepteurs des unités pédagogiques. Ils peuvent à partir d'un profil d'un apprenant particulier, sélectionner les LOs pertinents. Ils peuvent également à partir d'une description d'un objet d'apprentissage savoir les profils des apprenants s'y intéressant.

## 2.5 Principe du fonctionnement de SROA2

Notre système de recommandation SROA2 repose sur la structure d'une matrice appelée «matrice de pertinence» dont les lignes correspondent aux profils des apprenants et les colonnes correspondent aux descriptions des LOs. Les cellules de cette matrice contiennent la valeur « Vrai » ou « Faux » représentant la pertinence ou non de chaque objet d'apprentissage par rapport à chaque apprenant.

Le fonctionnement de notre système de recommandation de contenu relève de l'intelligence artificielle et plus particulièrement de la classification automatique. Ce système se base sur un classificateur construit en utilisant la technique SVM précédemment présentée. Nous avons alors utilisé SVM en vue de construire un modèle prédictif à partir d'un ensemble de couples formés par des entrées (des caractéristiques des apprenants et des LOs) et ses classes correspondantes (« pertinence » ou « non ») par l'estimation d'une fonction de prédiction qui décide la pertinence. Autrement dit, le but de l'apprentissage du classificateur SVM [10] [8] est d'estimer la fonction de pertinence exacte qui associe à chaque couple (profil / objet), une valeur (vrai / faux), en fonction de la convenance ou non de l'objet au profil. Mathématiquement, la fonction de pertinence se traduit par :  $F : P \times O \rightarrow \{ \text{Vrai}, \text{Faux} \}$ . Sachant que 'P' représente l'ensemble des profils des apprenants modélisés suivant IMS LIP [4] et 'O' représente l'ensemble des descriptions des LOs modélisés suivant LOM [3]. Pour chaque couple  $(p_i, o_j)$  appartenant à  $P \times O$ , la fonction de pertinence exacte F renvoie la valeur « vrai » si l'objet d'apprentissage 'o<sub>j</sub>' est pertinent pour l'apprenant 'p<sub>i</sub>' et la valeur « faux » sinon. La figure 3 montre le rôle du classificateur SVM :



**Fig. 3.** Le rôle du classificateur SVM.

## 2.6 Construction du classificateur SVM

### 2.6.1 Données utilisées pour construire le classificateur SVM.

- ❖ **Modèle d'apprenant.** Parmi l'ensemble des caractéristiques du modèle d'apprenant d'IMS LIP, nous retenons les attributs nécessaires à la personnalisation avec leurs valeurs possibles dans le tableau 1 ci-dessous :

**Tableau 1.** Caractéristiques d'apprenant nécessaires à la personnalisation.

Attribute	Value
Gender	Female {F}, men {M}
Language	Arabic {Ar}, French {Fr}, English {En} ...
Age	Low {L}, Medium {M}, High {H}
Education status	Elementary {E}, Junior {J}, Senior {S}, Undergraduate {U}, graduate {G}
Goal	Learn {L}, Search {S}, Contact {C}
Expertise fields	computer {PC}, groupware {Grp} ...
Expertise level	Low {L}, Medium {M}, High {H}, Expert {E}
Learning motivation	Low {L}, Medium {M}, High {H}
Cognitive style	Field dependence {FD}, field independence {FI}
Learning style	Doer {D}, Watcher {W}, Thinker {T}, Feeler {F}
Media preference	Audio {A}, Video {V}, Text {T}, Image {I}, Image and Text {IT}
Interactivity type	Reader {R}, Writer {W}, Larker {L}
Interactivity level	Very active {VA}, Active {A}, Visitor {V}, Inactive {I}
Disability	Visual {V}, Audio {A}, amputate member {M} ...
Interest field	E-Learning{EL}, Knowledge Management{KM}, Recommender system{RS}, Semantic Web{SW}, Web services {Ws}, Data Mining{DM},Artificial Intelligence {AI},...

- ❖ **Modèle d'objet d'apprentissage.** Egalement parmi l'ensemble des caractéristiques du modèle LOM, nous retenons les attributs nécessaires à la personnalisation avec leurs valeurs possibles dans le tableau 2 ci-dessous :

**Tableau 2.** Caractéristiques d'objet d'apprentissage nécessaires à la personnalisation.

Attribute	Value
Title	Recommender systems{RS}, Ontology in e-learning {ONTOEL}, Diagram of Nonaka and Takeuchi {DNONAKA}, Life cycle of a software{LC}, Artificial intelligence {AI},...
Language	Arabic {Ar}, French {Fr}, English {En}, ...
Keywords	Recommendation{REC}, ontology{ONTO}, knowledge management{KM}, Information System {IS},...
Format	Text{T}, Image{I}, Video{V}, Audio{Aud}, Application{App}, Model{Mod}, Multiformat{Mlt}...
Type	Paper{P}, Book{B}, Exercise{E}, Simulation{S}, Questionnaire{Q}, Diagram{D}, Figure{F}, graphic{G}, Index {I}...
Interactivity type	Active {A}, Descriptive {D}, Combine {C}
Interactivity level	Very Low {VL}, Low {L}, Medium {M}, High {H}, Very High {VH}
Semantic density	Very Low {VL}, Low {L}, Medium {M}, High {H}, Very High {VH}
Difficulty	Very easy {VE}, Easy {E}, Medium {M}, difficult {D}, Very Difficult {VD}
Learning duration	Year {Y}, Month {M}, Week {W}, Day {D}, Hour {H}, Minute {Mn}, Second {S}
User	Learner {L}, Teacher {T}, author {A}, Manager {M}
User language	Arabic {Ar}, French {Fr}, English {En}...
User age	Low {L}, Medium {M}, High {H}

- ❖ **Base de données des profils d'apprenants.** Selon le modèle d'apprenant que nous avons proposé au-dessus, nous avons collecté les informations relatives à des étudiants de l'université de Tlemcen, afin de constituer une base de données des profils des apprenants.
- ❖ **Base de données des descripteurs d'objets d'apprentissage.** Selon le modèle d'objet d'apprentissage que nous avons précisé au-dessus, nous avons décrit quelques ressources pédagogiques en informatique, afin de créer une base de données des descripteurs d'objets d'apprentissage.
- ❖ **Matrice de pertinence.** Nous rappelons que notre système de recommandation de contenu pédagogique repose sur la structure d'une matrice appelée « matrice de pertinence ». Elle contient un jeu de données construit à partir de la confrontation des jeux de données collectés qui sont relatifs aux apprenants et aux LOs.

### 2.6.2 Phases de construction du classificateur SVM. Il y a deux phases :

- *La phase d'apprentissage* ayant pour objectif d'ajuster correctement les paramètres nécessaires au meilleur fonctionnement du classificateur SVM, par exemple le choix du type de noyau (linéaire,

polynomial, gaussien,...) [8] et les paramètres du chaque noyau (exponentiel,...). Ceci pour induire une fonction de discrimination à partir du jeu d'entraînement.

- *La phase de test* permet d'évaluer les décisions du classificateur grâce au jeu du test.

**2.6.3 Implémentation du classificateur SVM avec Weka.** Nous avons utilisé l'algorithme du John Platt, SMO (Sequential Minimal Optimization) fourni dans l'environnement libre Weka en ajustant ses paramètres notamment le noyau qui est choisi de type « polynomial » avec un exponentiel égal à « 3 ». Ainsi, afin d'implémenter notre classificateur SVM suivant les deux phases, nous avons divisé notre jeu de données (qui est dans la matrice de pertinence) en deux sous ensembles, 70% pour l'apprentissage du classificateur et 30% restantes pour son test.

**2.6.4 Evaluation du classificateur SVM.** Après le test de notre classificateur SVM dans l'environnement Weka, nous avons obtenu les résultats suivants :

- *Temps pris pour construire le modèle prédictif* : 0,08 seconds ;
- *Matrice de confusion* :

Notre classificateur doit sélectionner les LOs qu'il estime être pertinents par rapport aux profils d'apprenants. Parmi les LOs sélectionnés, certains sont effectivement pertinents, d'autres ne le sont pas. On définit alors la matrice de confusion en tableau 3 :

**Tableau 3.** Matrice de confusion.

	LO sélectionné	LO non sélectionné
LO pertinent	VP = 11	FP = 1
LO non pertinent	FN = 0	VN = 3

On calcule alors le *taux de classification* ( $T_{class}$ ), la *Précision* ( $P$ ) et le *Rappel* ( $R$ ) d'un classificateur par :

$$T_{class} = (VP+VN) / (VP+VN+FN+FP); P = VP / (VP + FN); R = VP / (VP + FP).$$

La précision et le rappel sont liés puisqu'il est facile d'obtenir une précision élevée si le rappel est très faible et il est, de même, facile d'obtenir un rappel élevé au prix d'une précision très faible. Un classificateur doit donc être caractérisé à la fois par son rappel et sa précision. Pour faciliter les comparaisons entre différents classificateurs, on définit la mesure F qui prend en considération les deux indicateurs :  $F = 2 \times P \times R / P + R$ .

- *Evaluation des performances*:

Taux de classification	Précision	Rappel	Mesure F
93,3333 %	100 %	91,66 %	95,65 %

En effet, malgré que nous n'ayons pas utilisé des échantillons suffisamment nombreux, les résultats de notre expérience d'implémentation du classificateur SVM ont montré que les performances sont très prometteuses. Ainsi que l'efficacité de notre système de recommandation est meilleure par rapport aux autres approches telles que : l'agent recommandant [11] qui est basé sur le Web mining avec des règles d'association ; le système expert EPERS [9] ; PERSO [1] qui utilise LSA (Latent Semantic Analysis) et CBR (Case Based Reasoning) ; RACOFI [6] qui est basé sur le filtrage collaboratif et les règles d'inférences.

### 3 Conclusion

Dans cet article nous avons proposé une nouvelle démarche de construction d'un système de recommandation des LOs aux apprenants selon leurs profils. Ce système, qui peut être intégré dans un EIAH, permet aux concepteurs des unités pédagogiques d'explorer les LOs adéquats à un tel profil d'une part et de savoir quels sont les profils intéressés par un LO donné d'autre part. Nous avons proposé nos propres attributs des catégories d'informations de LOM et d'IMS LIP pour modéliser respectivement les LOs et les apprenants. Le fonctionnement de notre système est basé sur un classificateur SVM. D'après les résultats de son test, ses performances sont très prometteuses malgré que nous n'ayons pas utilisé des échantillons suffisamment nombreux.



Comme perspectives, nous espérons perfectionner notre classificateur SVM en intégrant la logique floue pour avoir des décisions beaucoup plus exactes et interprétables. Nous souhaitons également enrichir notre système de recommandation de contenu par l'élargissement de la recommandation aux scénarios d'apprentissage.

## Références

1. Chorfi, H et Jemni, M : PERSO: A System To Customize e-Training, Ecole Supérieure des Sciences et Techniques de Tunis 5, Tunisia, 2002.
2. Cornuéjols, A : Les SVM : Séparateurs à Vastes Marges, IIE & CNRS - Université de Paris-Sud, Orsay, France, 2002.
3. Final Draft Standard for Learning Object Metadata, Approved draft, Document IEEE 1484.12.1-2002, 2002, p 44. Available at: <http://ltsc.ieee.org/wg12/>
4. IMS Learner Information Package Specification, IMS Global Learning Consortium, 2001. Available at: <http://www.imspoint.org/profiles/index.html>
5. Koper, R : Modeling units of study from a pedagogical perspective – The pedagogical metamodel behind EML. Educational Technology, Expertise Centre, Open University of the Netherlands, 2002. Available at: <http://eml.ou.nl/introduction/articles.htm>
6. Lemire et al.: Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation, published in Technology and Smart Education, 2 (3). Troubador Publ., Leicester, UK, pp. 179-188. NRC 48535, August 2005.
7. Mobasher, B: Data Mining for Personalization. In The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, Kobsa, Nejdl (eds.). Lecture Notes in Computer Science, Vol.4321.Springer-Verlag, Berlin Heidelberg, 2006.
8. Mohamadally, H et Fomani, B : SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges, Versailles St Quentin, 16 janvier 2006.
9. Otair, M et al. : Expert Personalized E-learning Recommender System, University of Science and Technology Irbed, Jordan, 2003
10. Vapnik V., Statistical Learning Theory. John Wiley & Sons, 1998.
11. Zaiane, O: Building a Recommender Agent for e-Learning Systems, University of Alberta, Edmonton, Alberta, Canada, 2002.

## Architecture distribuée d'une plateforme du E-learning basée sur les Web services

Hakim Amrouche<sup>1</sup>, Amar Balla<sup>1</sup>, Djamel Charchal<sup>1</sup>, Adel Boukhedra<sup>1</sup>

<sup>1</sup>Institut National d'Informatique, BP 68 M, Oued-Smar, Alger, Algérie  
{ h\_amrouche, a\_balla, d\_charchal, a\_boukhedra }@ini.dz

**Résumé.** Plusieurs environnements d'apprentissage existent sur le Web, chacun de ces environnements offre différentes fonctionnalités (création de contenus, suivi des apprenants,...), ces derniers sont plus ou moins exploités. Les environnements actuels ont une architecture monolithique et ils se basent essentiellement sur les LMS et LCMS. Les acteurs de ces systèmes ont présenté des besoins de plus en plus croissants en termes de partage et de possibilité de réutilisation de données, ce qui a incité l'expansion de ces systèmes vers des environnements basés sur les systèmes répartis ou distribués. Dans cet article, nous allons présenter une architecture distribuée d'un environnement d'apprentissage basé sur les Web services. Les Web services de la plateforme sont distribués sur plusieurs sites. Un service peut être local ou externe à la plateforme. Un annuaire des différents services est mis en place afin de faciliter la recherche et l'invocation des services.

**Mots-clés:** E-learning, Web services, Architecture distribuée.

### 1 Introduction

La progression rapide des technologies de l'information et de la communication a donné naissance à une nouvelle forme d'apprentissage dite E-Learning. Les deux acteurs principaux dans ce mode d'apprentissage sont l'enseignant et l'apprenant. Pour que les enseignants et les apprenants puissent s'adapter à cette nouvelle méthode d'apprentissage, les plates-formes d'E-Learning doivent avoir les caractéristiques suivantes : la simplicité, l'interopérabilité des contenus pédagogiques, la réutilisation des contenus pédagogiques et l'adaptation des documents aux besoins des apprenants.

Plusieurs environnements d'apprentissages sont adoptés partout dans le monde. Ces systèmes permettent, par exemple, l'accès à des ressources pédagogiques, de procéder à des évaluations, et à la gestion des interactions entre enseignants et apprenants. Pour la mise en oeuvre de ces environnements, on a recours à l'utilisation des LMS et LCMS.

Les LMS (Learning Management System) sont des plates-formes développées pour faciliter aux enseignants la gestion des cours sur le Web. Ces systèmes permettent la diffusion des contenus pédagogiques, la gestion des apprenants (l'authentification, les inscriptions, leurs profils, leurs résultats et leurs progrès entre les différents modules de formation...). Les systèmes de présentation du contenu insistent sur la navigation, la visualisation et l'organisation des informations pour faciliter leur compréhension par les utilisateurs [4].

Les LCMS (Learning Content Management System) permettent aux créateurs de cours de créer, stocker, réutiliser, gérer et distribuer des contenus pédagogiques. Les LCMS favorisent l'accès et la recherche des données pédagogiques que celui de leur présentation. Les éléments d'information utilisés dans ces systèmes pour l'accès et la recherche sont des méta-données souvent standardisées [5].

Les environnements d'apprentissage actuels qui sont basés sur les LMS/LCMS ont une architecture monolithique (centralisée), c'est à dire une seule entité centralisée qui fournit toutes les fonctionnalités, ce qui rend leurs extensions difficiles [1]. En plus c'est difficile de les intégrer avec d'autres systèmes, de ce fait ces systèmes sont dit fermés [2].

Les acteurs de ces systèmes ont présenté des besoins de plus en plus croissants en termes de partage et de possibilité de réutilisation de données. Ce qui a incité l'expansion de ces systèmes vers des systèmes basés sur les environnements répartis et distribués [1][2]. Cette nouvelle génération est dite système d'apprentissage décentralisé. Ces systèmes sont basés, essentiellement, sur : le Web sémantique qui permet l'annotation des ressources pédagogiques, c'est-à-dire, il permet d'attribuer plus de sémantique à ces ressources ; ou sur les ontologies qui facilitent la représentation et l'organisation des contenus

pédagogiques dans le but de permettre le partage et la réutilisation de ces contenus, ou bien sur les Web services qui assurent une bonne communication entre les différentes applications.

Dans cet article, nous nous intéressons à l'utilisation des Web services pour la construction d'un environnement d'apprentissage distribué. La suite de l'article est organisée comme suit : dans la section 2, nous allons donner une présentation des Web services. Par la suite, dans la section 3, nous allons présenter les détails de l'architecture de la plateforme proposée. Dans la section 4, quelques détails d'implémentation sont donnés et nous terminons par une conclusion et des perspectives.

## 2 Les Web Services

Les Web services ont pour objectifs principales l'interopérabilité entre les applications via le Web indépendamment des plates-formes (langages de programmations, systèmes d'exploitation,...). Ils facilitent l'accès aux applications et les échanges de données. Des exemples de services actuellement disponibles concernent les prévisions météorologiques, le E-commerce et le E-learning[3][8].

La notion de «Web service» désigne essentiellement une application mise à la disposition des utilisateurs sur Internet par un fournisseur de service. Un web service peut être accessible depuis une autre application (un client, un serveur ou un autre Web service) à travers le Web, et ceci en utilisant les protocoles de transports disponibles. Ce service applicatif peut être implémenté comme une application autonome ou comme un ensemble d'applications [6][7].

### 2.1 Architecture des Web Services

L'architecture de référence comporte les trois éléments suivants [7]:

- **Le fournisseur de service** : c'est le propriétaire du service.
- **Le client** : c'est le demandeur de service. Il est constitué par l'application qui va rechercher et invoquer un service.
- **L'annuaire des services** : c'est un registre qui contient les descriptions de services.

Ces trois éléments de l'architecture interagissent entre eux selon trois types d'opérations : les opérations de publication, de recherche et de liens d'opérations. La figure suivante montre l'architecture de référence des web services.

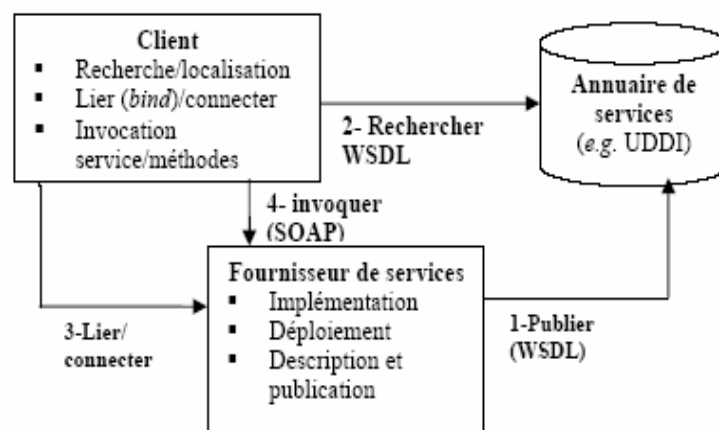


Fig. 2. Architecture des Web Services [7]

Le fournisseur de services définit la description WSDL (Web Service Description Language) de son service. WSDL est un langage qui décrit, au format XML, les méthodes pouvant être invoquées et le point d'accès (URL, port, etc..) du Web service. Les descriptions WSDL sont publiées dans un annuaire de service UDDI (Universal Description Discovery and Integration) qui peut être public ou privé. Le standard UDDI vise à décrire une manière standard, de publier et d'interroger les Web services. Le client

utilise les facilités de recherche disponibles au niveau de l'annuaire pour retrouver et sélectionner un service donné. Il récupère ensuite les informations nécessaires, sous format WSDL, lui permettant de se connecter au fournisseur du service et d'interagir avec le service considéré.

La communication entre le client et le fournisseur est assurée par le protocole SOAP (Simple Object Access Protocol). SOAP est un protocole d'échange inter applications indépendant de toute plate-forme, basé sur le langage XML.

Le mécanisme de découverte offert par l'UDDI se base essentiellement sur la recherche par mots clés, ceci est dû au manque d'informations sémantiques dans le fichier WSDL. Pour surmonter ces problèmes une nouvelle génération de service dite web services sémantiques a été proposée.

### **3 Architecture du système proposé**

L'objectif de notre travail est de proposer une architecture distribuée et flexible d'un environnement d'apprentissage. L'architecture proposée est constituée d'un ensemble de Web services qui peuvent être distribués sur le Web. Notre plateforme doit avoir les mêmes fonctionnalités qu'une plateforme classique du E-learning. En plus, elle doit être distribuée d'une manière transparente aux acteurs de la plateforme. Alors, lors de notre conception, nous avons pris en considération toutes ces fonctionnalités.

Dans notre architecture chaque fonction de la plateforme du E-learning est représentée sous forme d'un Web service. Il y a des fonctions qui nécessitent plus d'un Web service, il s'agit de Web service composite qui est défini comme étant la composition de plusieurs services.

Dans le cas pratique on peut trouver deux Web services qui offrent la même fonctionnalité, par exemple deux services pour la création d'un cours, le premier permet d'intégrer des animations et l'autre non. C'est pour ces raisons que nous avons autorisé d'intégrer des services externes dans notre plateforme, aussi c'est possible d'utiliser nos propres services par d'autre plateforme de formation, Ceci afin de rendre notre plateforme ouverte. Un service externe peut être utilisé par un acteur de notre plateforme pour lui assurer une certaine fonctionnalité dans le cas où notre service local ne satisfait pas ces besoins.

Nos services ne sont pas tous accessibles par les autres plates-formes, nous distinguons deux catégories de services, à savoir les services locaux et les services publics. Les services locaux sont accessibles uniquement par les utilisateurs internes de la plateforme. Par contre, les autres services sont accessibles pour les d'autres services et d'autres environnements.

Pour simplifier l'opération de l'utilisation (recherche et invocation des services Web) des services externe nous avons opté pour l'implémentation d'un registre UDDI propre à notre plateforme. Ce registre contient la description des Web services public et externe à la plateforme.

La figure 2 présente l'architecture globale de la plateforme.

#### **3.1 Les acteurs de la plateforme**

Nous avons retenu trois catégories d'acteurs dans notre plateforme à savoir :

- **L'administrateur:** Il s'occupe de l'inscription des acteurs dans la plateforme. Il a l'accès aux profils utilisateurs et à la gestion de tous les acteurs. Il est responsable de la gestion de l'annuaire des services.
- **L'enseignant:** Il se charge de la création des cours et des évaluations ainsi que le suivi des apprenants.
- **L'apprenant:** Il représente l'acteur ayant été intégré dans la plateforme dans le but est de suivre une formation donnée.

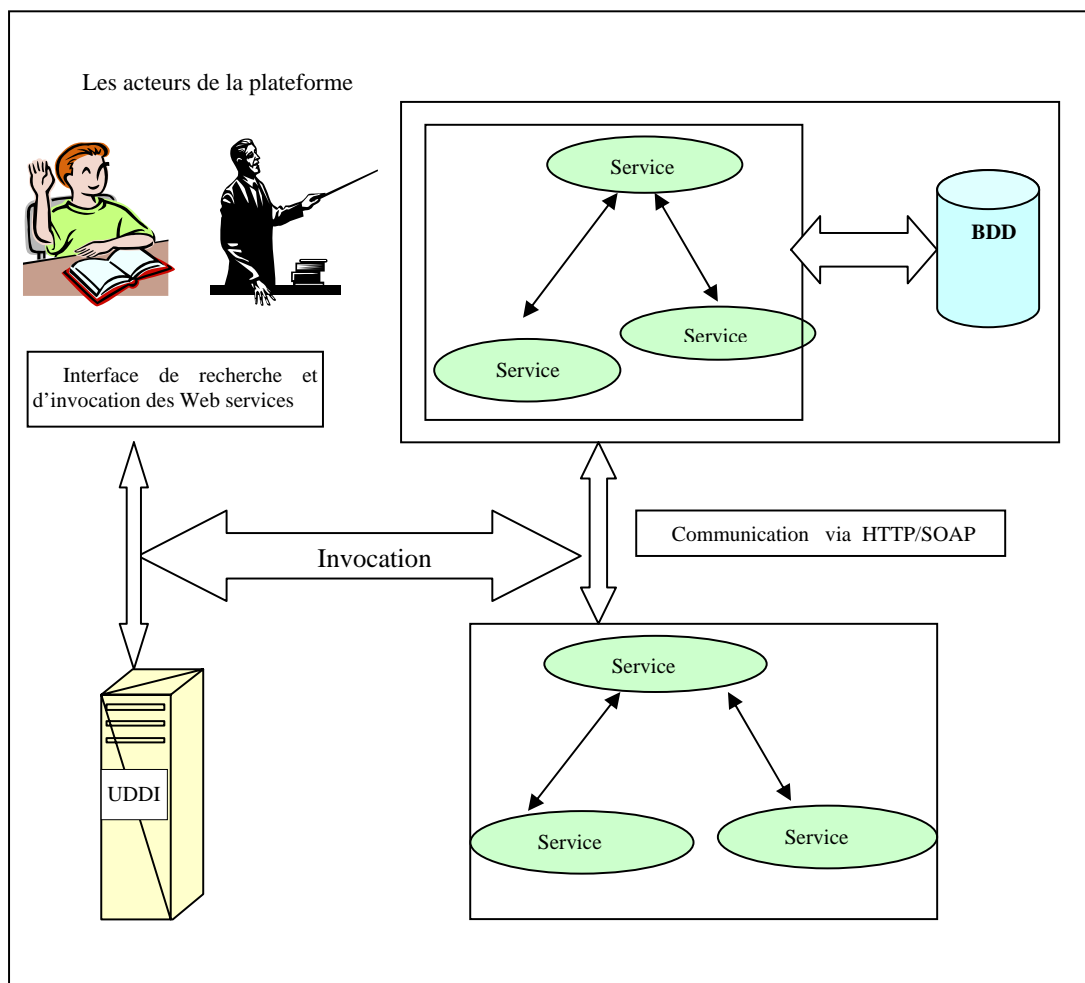


Fig. 2. Architecture du système proposé

### 3.2 Les services de la plateforme

Les services qui constituent notre plateforme sont présentés comme suit :

- **Le service de gestion des utilisateurs:** C'est un service qui s'occupe de l'inscription des nouveaux apprenants ou enseignants, la modification des informations d'un acteur ou de sa suppression à partir de la plateforme. Ce service prend en charge aussi la vérification de l'identité des acteurs, et ceci afin d'autoriser ou de refuser les accès aux ressources de la plateforme.
- **Le service de gestion des profils:** C'est un service important car il est très utilisé par les autres services. Ce service garde les informations de profil des apprenants et l'historique de leurs activités au sein de la plateforme. Ce service gère une base de données des profils des apprenants.
- **Le service de consultation :** L'apprenant peut consulter son profil à travers ce service. L'enseignant peut aussi consulter les profils des apprenants pour voir leurs états d'avancement.
- **Le service de création des cours:** C'est un service que l'enseignant peut utiliser pour créer des cours et des épreuves pour les apprenants. C'est un service assez complexe, dans notre travail, nous ne sommes pas intéressés au développement de ce service. Un travail intéressant sur le service de création des cours est présenté par Ivan Madjaro dans [8].

- **Le service de suivi des apprenants:** Ce service se charge d'effectuer des évaluations pour les apprenants. Il reçoit les épreuves et les exercices réalisés par les apprenants. Par la suite, il va effectuer les évaluations nécessaires selon une certaine méthode d'évaluation.
- **Le service consultation des cours et épreuves:** C'est un service qui permet de récupérer les cours et les épreuves afin de les afficher aux apprenants.
- **Le service d'interrogation:** Le rôle de ce service est de répondre aux requêtes des utilisateurs pour la recherche des ressources pédagogiques sur le Web (entrepôts distants ou à partir de la base locale).

### **3.3. Interactions entre les services de la plateforme proposée**

Les services de la plateforme ne sont pas indépendants, il y a une forte interaction entre ces services. Dans ce qui suit, nous allons détailler les principales interactions entre les services:

1. **Interaction entre le service de gestion des utilisateurs et le service de gestion des profils :** Les deux services se communiquent les informations qui concernent les apprenants lorsqu'un nouveau apprenant est inscrit dans la plateforme, le service de gestion des utilisateurs communique les informations nécessaires au service de gestion des profils afin qu'il puisse créer un profil vierge pour l'apprenant en question. Le service de gestion des utilisateurs communique aussi avec le service de la gestion des profils pour vérifier l'identité de l'acteur lors de sa connexion à la plateforme.
2. **Interaction entre le service de consultation et le service de gestion des profils :** Lorsqu'un apprenant veut visualiser son état d'avancement et son historique, alors il fait appel au service de consultation, ce dernier va se charger de contacter le service de gestion du profil. La même opération est réalisée par l'enseignant s'il veut visualiser les profils des apprenants.
3. **Interaction entre le service de suivi et le service de gestion des profils :** Lorsque le service de suivi termine l'évaluation des exercices et des épreuves réalisés par l'apprenant alors, il contacte le service du profil pour lui communiquer les résultats afin de mettre à jours le profil de l'apprenant.
4. **Interaction entre le service de consultation des cours et le service de gestion des profils :** Les deux services communiquent pour garder l'historique des apprenants qui ont consulté des cours. Le service de consultation utilise les informations sauvegardées dans le profil apprenants afin d'adapter le cours à leur profils.
5. **Interaction entre le service de suivi des apprenants et le service de consultation des cours et des épreuves :** Le service de suivi reçoit du service de consultation des cours et des épreuves les exercices et les examens effectués par l'apprenant, afin de les évaluer et de mettre à jour l'état d'avancement des apprenants.
6. **Interaction entre le service d'interrogation et le service de gestion des profils :** Le service d'interrogation récupère les préférences de l'apprenant à partir du service de gestion des profils en vu de répondre aux requêtes de recherche. Les informations sur le profil de l'apprenant sont utilisées pour enrichir la requête de recherche.
7. **Interaction entre le service d'interrogation et le service de création des cours :** Le service de création de cours fait appel au service d'interrogation afin de lui fournir des objets pédagogique. Cela permet au concepteur du cours d'utiliser/réutiliser des objets pédagogiques déjà existants.

Nous remarquons que le noyau de ces interactions est le service de gestion des profils des apprenants. Pour cela ce service est propre à notre plateforme, il gère la base des profils des apprenants. Afin d'assurer l'interopérabilité avec les autres services nous avons opté pour utiliser XML pour la représentation des profils. Ainsi le profil de chaque apprenant est représenté sous forme d'un fichier XML.

### **3.4 L'annuaire des services**

Cet annuaire contient la description WSDL des Web services interne et externe à la plateforme. Cette UDDI est gérée par l'administrateur de la plateforme. C'est lui qui prend en charge la publication des nouveaux services et de chercher des descriptions de Web services externe afin de les intégrer dans la plateforme. Une interface est offerte pour les acteurs de la plateforme pour faire des recherches des Web services afin de les invoquer. La recherche se base essentiellement sur des mots clé dans l'annuaire UDDI.

## **4 Conclusion et perspectives**

Nous avons présenté dans cet article une architecture ouverte, distribuée et flexible d'un environnement d'apprentissage à base de Web services. Cette architecture est composée d'un ensemble de services distribués sur le Web. Les services de notre plateforme peuvent être utilisés par d'autres environnements, en plus c'est possible d'intégrer d'autres services à partir d'autres environnements. La communication entre les services est réalisée par un protocole de haut niveau afin d'assurer l'interopérabilité des services. L'implémentation d'un annuaire UDDI permet l'utilisation des services externes à la plateforme.

Comme perspectives nous envisageons d'implémenter la totalité des services de la plateforme en premier lieu, par la suite d'automatiser la procédure de découverte et de sélection des Web services.

## **References**

1. Brusilovsky P. (2004). KnowledgeTree: A distributed architecture for adaptive e-learning. In: Proceedings of The Thirteenth International World Wide Web Conference, WWW 2004 New York, NY, 17-22 May, 2004, ACM Press, pp. 104-113.
2. Dolog P., Henze N., Nejdl W., Sintek M. ( 2004). Personalization in Distributed eLearning Environments. In Proc. of WWW2004 - The Thirteen International World Wide Web Conference, New York, USA, 2004. ACM.
3. Vossen G., Westerkamp P. (2003). E-Learning as a Web Service Seventh International Database Engineering and Applications Symposium (IDEAS'03)
4. Dehors S., Giboin A., Faron-Zucker C., Stromboni J.-P. (2005). Des Annotations Sémantiques pour Apprendre : l'Expérimentation QBLS. In Actes de la Journée thématique WebLearn sur le Web sémantique pour le e-Learning, plateforme AFIA'2005, Nice.
5. Duval E., Forte E., Cardinaels K., Verhoeven B., Van Durm R., Hendrikx K., Wentland-Forte M., Ebel N., Macowicz M., Warkentyne K., Haenni F. (2001). The ARIADNE Knowledge Pool System. In Communications of the ACM, 44 (5), pp. 73-78.
6. Hubert K, Monfort V. (2003). Les Web services Edition DUNOD 2003
7. Kellert P., Toumani F. (2004). Les Web services sémantiques . Revue I3 (Information - Interaction – Intelligence) . Volume Hors Série 2004.
8. Ivan Madjarov .(2005). Des services web pour le e-Learning. e-TI - la revue électronique des technologies d'information, Premier Numéro, 28 octobre 2005.



## Architecture d'un entrepôt pédagogique et sa modélisation sémantique

Nawel ILES, Azeddine CHIKH, Sidi Mohammed CHOUTI

Faculté des sciences de l'ingénieur  
Université de Tlemcen  
Chetouane, BP 230, 13000-Tlemcen  
{n\_iles, az\_chikh,choutism}@univ-tlemcen.dz

**Résumé.** Les nouvelles Technologies de l'Information et de la Communication "TIC" améliorent profondément nos façons de nous informer, de communiquer et de nous former. Une masse importante de ressources pédagogiques est produite à travers de nombreuses universités. La ré-utilisation des objets pédagogiques ainsi produits en local est faible et même inexistante entre les universités. Ceci peut s'expliquer en particulier par le manque de connaissance de la part des enseignants sur les ressources existantes. Nous proposons dans ce travail de recherche un modèle d'entrepôt distribué qui repose sur la description des OPs par les métadonnées de LOM et les concepts des ontologies de domaine. Ces deux descriptions complémentaires permettent l'interrogation sémantique des OP.

**Mots-clés:** E-learning, objet pédagogique, entrepôt, métadonnée LOM, Ontologies, Partage, ré-utilisation

### 1 Introduction et Problématique

Le E-learning est défini comme l'utilisation des nouvelles technologies multimédias et de l'internet, pour améliorer la qualité de l'apprentissage en facilitant l'accès à des ressources et des services, ainsi que des échanges et la collaboration à distance. Les objets pédagogiques « OP » constituent l'ensemble des informations, parties de cours, de programmes, de thèses, etc. qui permettent de véhiculer et de transmettre des concepts et contenus d'enseignements. Un OP numérique peut être de différents formats : .PDF, .DOC, .JPG, etc. Les OP sont créés par différentes universités et sont essentiellement utilisés au sein de l'établissement qui les a créés. Ainsi, une université n'a pas accès aux OP des autres universités (sauf relations informelles entre enseignants). Il en résulte un manque de capitalisation des savoirs faire et des objets ; Compte tenu du coût de production de tels OP et de l'expertise nécessaire pour les produire, il est primordial de les rendre facilement accessibles, exploitables et réutilisables.

Notre proposition vise à mettre à disposition les OP existants à un ensemble d'universités organisées en réseau. Cette mise à disposition des OP n'est pas suffisante pour garantir leur réutilisation. Elle doit s'accompagner de moyens de recherche et d'accès à l'information utile. La description LOM qui peut être associée à chaque OP constitue une première description indispensable à cet accès. Cependant elle n'encadre pas suffisamment les choix faits lors de la description des contenus des OP. A l'heure du Web sémantique, il est alors nécessaire d'associer des annotations sémantiques au contenu des OP pour permettre de les retrouver. Ces deux descriptions complémentaires permettent l'interrogation sémantique des OP. Les métadonnées et les annotations de chaque entrepôt local sont dupliquées dans un méta entrepôt central et partagé, accessible à partir d'un serveur commun. La suite de l'article est organisée comme suit. Dans la section 2, nous présentons l'état de l'art ; nous y abordons en particulier l'utilisation des entrepôts et celle des ontologies. Dans la section 3, nous proposons l'architecture de l'entrepôt pédagogique et le modèle sous-jacent. Nous terminons l'article par une conclusion et les perspectives aux travaux présentés.

## 2 Etat de l'art

### 2.1 E-Learning: objets pédagogiques et norms

#### 2.1.1 Objet pédagogique

En E-Learning, la forme fondamentale d'une ressource pédagogique est l'« objet pédagogique » (OP); Le groupe de travail des IEEE-LTSC a proposé une définition d'un OP qui est la suivante : « Un objet pédagogique est défini comme toute entité numérique ou non qui peut être utilisée, réutilisée ou référencée pendant des activités d'apprentissage assistées par ordinateur (enseignement intelligent assisté par ordinateur, environnements d'enseignement interactifs, systèmes d'enseignement à distance, environnements d'apprentissage collaboratif) ».

#### 2.1.2 Les métadonnées et Ontologies

Afin de faciliter l'accessibilité, la recherche, le partage et la ré-utilisation, chaque OP doit être décrit à l'aide de métadonnées et d'annotations sémantiques. En effet, un OP non indexé ou annoté ne pourra pas être retrouvé et à fortiori réutilisé. Cette étape d'indexation est donc indispensable, elle a été abordée dans différents travaux présentés dans les paragraphes qui suivent.

**a- Les métadonnées** Les métadonnées peuvent être définies comme étant des données relatives à d'autres données. Dans le domaine de l'apprentissage automatique, les métadonnées servent à décrire les documents pédagogiques afin de les rendre plus facilement identifiables (accessibles) et plus manipulables (interopérables, réutilisables, durables, adaptables). Ces métadonnées sont structurées suivant des catégories ou champs sémantiques « descripteurs ». Chaque champ représente une caractéristique particulière sur la ressource comme par exemple son titre [2]. Dans notre travail, nous appuyons sur LOM qui permet une description à la fois fine et précise.

#### Learning Object Metadata « LOM »

Le standard LOM spécifie la syntaxe et la sémantique des méta-données décrivant des ressources pédagogiques numériques ou non et définit les attributs nécessaires à une description complète des ressources pédagogiques à partir des 78 attributs divisés en neuf catégories : Généralité, Cycle de vie, Méta-métadonnées, Technique, Pédagogie, Droits, Relation, Annotation, Classification.. LOM est très complet et permet d'identifier tout type de contenu, et plus spécifiquement les contenus éducatifs. Les profils d'applications ont été introduits afin de répondre aux besoins spécifiques et concrets des utilisateurs. Un profil d'application est une instance d'un modèle comme LOM dans un contexte particulier. De fait, cela signifie interpréter, raffiner, étendre ou parfois même simplifier les syntaxes et les sémantiques des méta-données LOM.

#### b- Les Ontologies

LOM n'est pas assez puissant pour assurer le partage des OP et leur réutilisation. En effet, la sémantique des contenus est mal représentée puisqu'elle se limite souvent à des mots-clés. Afin de la compléter, nous suggérerons l'utilisation d'ontologies de domaine. En effet, une ontologie regroupe les concepts qui représentent l'ensemble des connaissances d'un domaine en une spécification explicite et formelle. Elle montre les relations ainsi que les règles d'associations qui existent entre ces concepts. Ainsi, une ontologie peut fournir le vocabulaire et sa sémantique, pour exploiter le contenu des métadonnées associées aux objets pédagogiques annotés [8]. La description d'une ontologie repose sur différentes normes : RDF (Resource Description Framework), OWL (Web Ontology Language), etc. Différents travaux de la littérature s'intéressent à l'indexation ou l'annotation sémantique, citons les projets suivants : TRIAL-SOLUTION [4], Memora (MEMoire Organisationnelle Appliquée à l'apprentissage en ligne) [1], Arpem (Archivage et Ressources PEdagogiques Multimédia) et le projet Ariadne [5]. Les différents travaux cités se sont intéressés à la construction des entrepôts pédagogiques en se basant sur différentes façons d'annoter les OP et sur différentes approches: les projets Ariadne, Arpem sur les métadonnées, Memora, Trial Solution et karina sur les ontologies. Ce type de modèle permet à chaque université de construire son propre entrepôt pédagogique mais leurs contenus restent réservés toujours qu'à un public très limité. Pour faire face à cette contrainte et élargir la diffusion des OP, notre approche consiste à envisager un système qui permette, d'un côté de capitaliser les OP, les indexer afin de les rendre ré-utilisables et d'un autre côté qui permette à chaque OP annoté d'être accessible et exploité

par les acteurs de différentes universités. Ainsi, nous proposons de construire un entrepôt pédagogique dit « serveur », qui capitalise l'ensemble des OP annotés en un seul endroit.

## 2.2 Les entrepôts pédagogiques en e-learning

De nombreuses définitions d'un entrepôt ont été proposées, nous retenons celle de Inmon-William H. (Inmon, 2002) « un entrepôt de données est une collection de données orientées sujet, intégrées, filtrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision ». Les données de l'entrepôt doivent avoir les caractéristiques suivantes :

- Orientées sujet : les données doivent être classées par sujet ou par domaine ;
- Intégrées : les données sont le résultat de l'intégration de données en provenance de sources différentes et hétérogènes ;
- Historisées : l'entrepôt doit archiver les données dans leurs différentes versions ;
- Filtrées : l'entrepôt ne doit contenir que des données pertinentes.
- Non volatile : les données sont stockées dans l'entrepôt et elles ne peuvent pas être modifiées.

Les collections des OP sont régies par les mêmes caractéristiques. Chaque OP doit être sauvegardé de manière permanente, présenté sous différentes versions et peut être issu de différentes sources. Ainsi, nous proposons un entrepôt pédagogique qui permet de capitaliser non seulement les OP, mais aussi, toute sorte d'information liée à l'OP et qui peut être utile à l'utilisateur.

## 3 Architecture et modélisation de l'entrepôt pédagogique distribué

### 3.1 Architecture d'un entrepôt pédagogique local

Le contenu de l'entrepôt pédagogique associé à une université, doit être constitué d'un ensemble de ressources pédagogiques de différents formats (pdf, doc, jpg, etc.) et sur différents supports (papier, électronique, multimédia, etc.). Une ressource pédagogique est composée d'un ensemble d'objets pédagogiques « OP » reliés entre eux. Chaque OP est décrit par les métadonnées LOM afin de le retrouver et de le ré-utiliser. L'entrepôt pédagogique associé à une université sera composé de deux parties : une base de données classique (relationnelles, orientées objets ou relationnelles-objets) qui va contenir l'ensemble des OP filtrées et un ensemble de descripteurs LOM et de descripteurs sémantiques résultant après la phase d'indexation. De cette façon, on sépare le stockage des OP de ses descripteurs tout en gardant un lien entre eux, et on facilite la recherche des OP en se basant sur les métadonnées LOM, et les annotations sémantiques. Notre entrepôt local est représenté par l'architecture suivante :

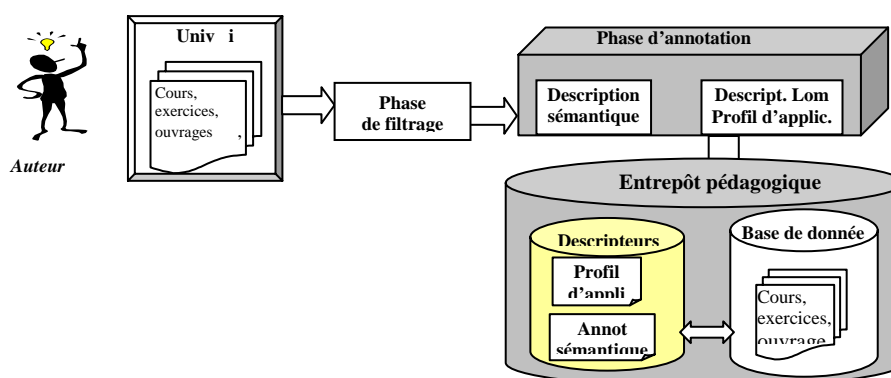


Fig. 1. Modélisation d'un entrepôt local

L'intégration d'un OP dans l'entrepôt distribué se fait en deux phases :

**1- Phase de filtrage :** il s'agit de sélectionner les OP pertinents pour les capitaliser, tout en gardant les relations entre les objets appartenant à la même ressource.

**2- Phase de qualification :** A chaque OP sont associés un ou des descripteur (s) LOM et une ou des annotation(s) sémantiques par rapport aux ontologies de domaine[9]. La mise en relation entre les ontologies et les contenus des OP s'appuie sur les techniques d'indexation automatique telles que abordées dans [6]

### 3.2 Qualification d'un Objet pédagogique

Une description LOM est rattachée à chaque objet pédagogique (qu'il soit élémentaire ou composé). Les méta-données utiles pour l'application donnée peuvent donc ensuite être filtrées via le profil d'application, en fonction de celle-ci. Une modélisation de cette description est montrée dans la figure 2.

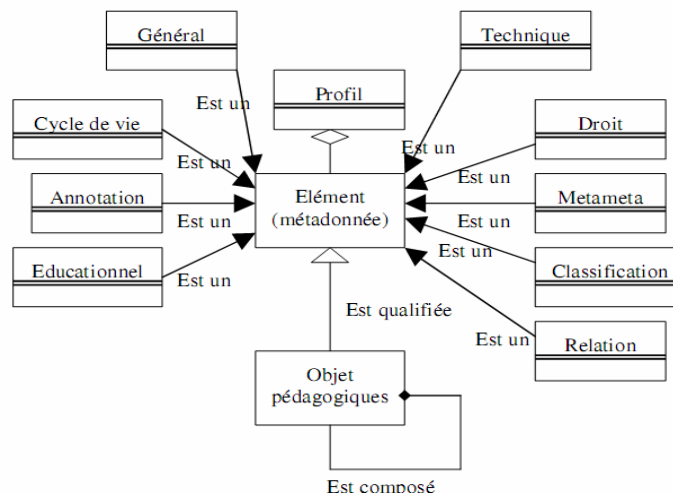


Fig. 2. Description LOM d'un OP

Les méta-données associées permettent de renseigner, d'une manière bien classifiée, les différentes informations nécessaires sur chaque objet d'apprentissage, de façon à ce que les recherches ultérieures soient rendues plus efficaces.

Afin d'assurer la réutilisation d'un OP, la représentation sémantique est basée sur la notion d'ontologie. Nous utilisons une ontologie de thème (Hernandez, 2005) de la formation qui rassemble les thèmes, notions et connaissances à appréhender. Cette ontologie décrit l'ensemble des notions en lien avec le domaine et les représentent à partir de leur lien sémantique. Par exemple, dans le domaine de l'informatique et plus précisément des bases de données, la notion (concept) de « base de données relationnelles » se « conceptualise » à partir d'un « modèle Entité-Association ». L'apprentissage d'une notion donnée pouvant demander un certain nombre de pré-requis, les notions correspondant à des pré-requis d'une autre notion sont également représentées dans l'ontologie (par exemple la notion d'« attribut » doit être assimilée pour appréhender la notion de « dépendance fonctionnelle »). La figure 3 montre un modèle de description d'ontologie de thème.

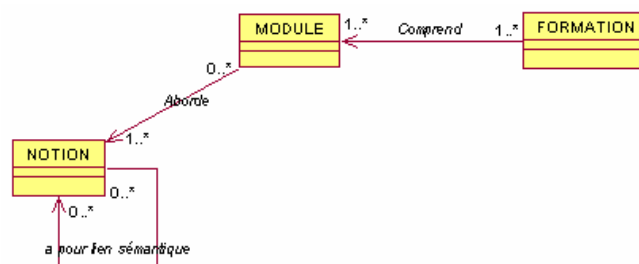


Fig. 3. Modèle de description thématique

### 3.3. Architecture générale du méta entrepôt pédagogique

Dans le cadre universitaire, il existe différentes sources des OP: supports de cours, exercices, thèses, rapports, etc..., disponibles par les utilisateurs au niveau de chaque université. Cependant, il leur est impossible d'accéder de façon efficace, et de trouver les OP pertinentes adaptés à leurs besoins. D'un autre côté, ces OP restent localement réservés, et non accessibles par les utilisateurs d'autres d'universités. Dans notre approche, nous proposons une architecture distribuée qui consiste à mettre en place un système qui va permettre de :

- Capitaliser au niveau de chaque université, tous les OP pertinents produits ;
- Annoter chaque OP afin de la rendre accessible et ré-utilisable. Cette annotation se base sur une description par les métadonnées LOM et par les annotations sémantiques ;
- Créer un méta-entrepôt pédagogique qui sera alimenté par l'ensemble des descripteurs des entrepôts des universités. Ainsi chaque ressource pourra être partagée entre universités, et elle sera accessible à distance (tout en respectant les rassemblements dans un endroit unique).

L'architecture générale du modèle est la suivante :

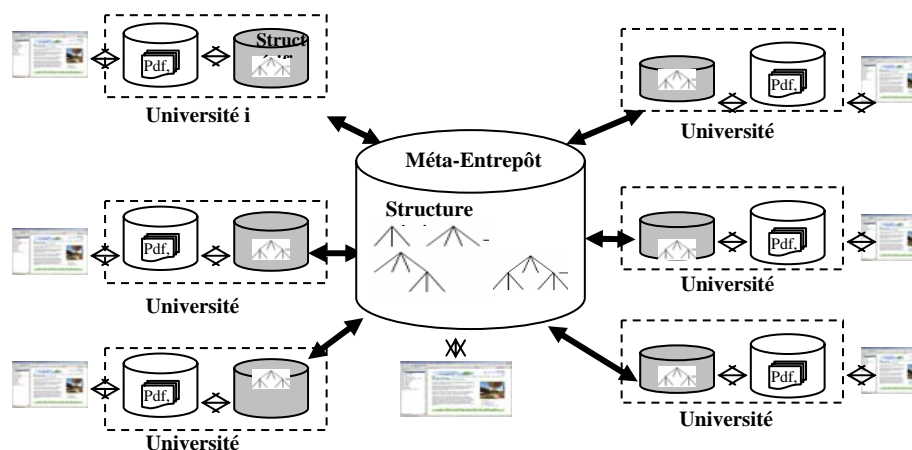


Fig. 4. Architecture distribuée de l'entrepôt pédagogique

Les entrepôts pédagogiques se basent sur deux types de structures : des structures spécifiques et une structure générique. La structure spécifique caractérise l'organisation d'un OP. Chaque entrepôt pédagogique peut contenir plusieurs structures logiques spécifiques. Elle permet la description sémantique d'un OP, et la structure générique représente une collection de représentation spécifique. Les entrepôts pédagogiques ayant des structures similaires ou identiques sont regroupés dans une même structure spécifique afin de faciliter la recherche.

## 4 Conclusion

Pour conclure, nous avons proposé un modèle d'entrepôt distribué qui permet l'indexation des objets pédagogiques, leur recherche et leur ré-utilisation de façon à répondre aux besoins des apprenants ou des enseignants. L'originalité de notre approche peut être résumée par le fait qu'elle :

- s'appuie sur un enrichissement des descriptions LOM par des annotations sémantiques par rapport aux ontologies de domaine ;
- repose sur une architecture distribuée où les spécificités de chaque université sont prises en compte ;
- permet une interrogation homogène des objets existants via une méta description des objets ;
- assure des possibilités avancées de traitement via l'utilisation des fonctionnalités d'un entrepôt de données.

En vue de pouvoir retrouver les objets pédagogiques directement lors de l'implémentation d'un nouveau cours en ligne sur un LMS, nous comptons l'interfacer avec l'entrepôt distribué.

## **Références**

1. Abel, M.-H., Lenne, D., Moulin, C., & Benayache A.: Gestion des ressources pédagogiques d'une e-formation [Pedagogical resources for e-learning]. Document Numérique, 7(1-2) (2003) 111-128.
2. Bourda, Y.: Objets pédagogiques, vous avez dit objets pédagogiques?, Supélec, Plateau de Moulon, F91192, Gif-sur-Yvette, CEDEX (2001)
3. Bouzeghoub, A., Defude, B., Duitama, J.F., Lecocq, C.: Un modèle de description sémantique de ressources pédagogiques, basé sur une ontologie de domaine. Sticf, Volume 12 (2005)
4. Buffa, M., Dehors, S., Faron-Zucker, C., Sander, P. : Vers une approche sémantique dans la conception d'un système d'apprentissage. Revue du projet TRIAL-SOLUTION, Plate forme AFIA nice (2005)
5. Chikh, A.: Une approche méthodologique de réutilisation en ingénierie de document. Thèse de Doctorat, Alger: INI (2003)
6. Hernandez, N.: Ontologies de domaine pour la modélisation du contexte en Recherche d'Information. Thèse de doctorat, Toulouse: Université Paul Sabatier (2005)
7. Hernandez, N., Mothe, J., Ralalason, B., Ramamonjisoa, B., Stolf, P.: A Model to Represent the Facets of Learning Objects. Interdisciplinary Journal of Knowledge and Learning Objects, Informing Science Institute, Volume 4 (à paraître), Santa Rosa-USA (2008)
8. Inmon, W.-H.: Building the data warehouse. New York (2002)
9. Snae, C. & Brueckner, M.: Ontology-Driven E-Learning System Based on Roles and Activities for The Learning Environment. Interdisciplinary Journal of Knowledge and Learning Objects, Volume 3, (2007) 1-17
10. Varlamis, I., & Apostolakis, I.: The Present and Future of Standards for E-Learning Technologies. Interdisciplinary Journal of Knowledge and Learning Objects, Volume 2, (2006) 59-76.

## **Références sur le WEB.**

1. Boutemedjet, S.: Web Sémantique et e-Learning. [Semantic web and e-learning], Cours IFT6261, [http://www.iro.umontreal.ca/~aimeur/cours/ift6261/Presentations\\_etudiants/WebSemantiqueE-learning.pdf](http://www.iro.umontreal.ca/~aimeur/cours/ift6261/Presentations_etudiants/WebSemantiqueE-learning.pdf) (2004)
2. Gasevic, D., et Hatala, M.: Searching context relevant learning resource using ontology mappings, International Workshop on Applications of Semantic Web Technologies for E-learning (SW-EL), Winston-Salem State University, <http://www.win.tue.nl/SW-EL/2005/swel05-kcap05/proceedings> (2005)
3. Knightm C., Gasevicm D., & Richardsm G.: Ontologies to integrate learning design and learning content. Journal of Interactive Media in Education, (07), ISSN:1365-893X. <http://www-jime.open.ac.uk/2005/07/> (2005)

## *Web Services*

---



## Bringing semantics to web services: Model driven approach

Djamel Amar Bensaber<sup>1</sup>, Djamel Benslimane<sup>2</sup>, Mimoun Malki<sup>1</sup>

<sup>1</sup> EEDIS laboratory University of Sidi Bel Abbes , Algeria

<sup>2</sup> LIRIS laboratory UCBL LYON FRANCE

{amarbensaber,mymalki}@gmail.com, djamel.benslimane@liris.cnrs.fr

**Abstract.** The semantic web promises automates invocation, discovery and composition of web services by enhancing services with semantic descriptions. This paper describes a model driven approach to facilitate the construction of OWL-S specifications. The methodology is divided into three main steps. In the first step we reverse engineered WSDL documents into UML profile models that enable the use of high-level graphical models as an integration platform for semantic web services. In the second step, suitable domain ontologies are used for the semantic annotation of the UML models. Finally, in the third step a conversion tool will generate automatically the OWL-S description from these UML models. The UML profile provides flexibility as it can express multiple semantic web service concepts.

**Keywords:** WSDL, UML, MDA, Reverse engineering, Semantic web service.

### 1 Introduction

Web services are gaining momentum as a new breed of distributed Web applications. Based on standard internet protocols such as SOAP, WSDL [1], and UDDI, Web services have become a key technology of B2B integration and service-oriented architecture. However, as the number of Web services increases and different people may describe similar or identical things differently, how to automatically process services at runtime is becoming an important issue. To make Web services more accessible to automated processes, there is a growing interest in bringing semantics into Web services.

As a *de facto* standard, OWL-S [2] defines a core set of markup language constructs for describing the properties and capabilities of Web services in unambiguous, computer-interpretable forms. OWL-S facilitates the automation of Web service tasks including discovery, execution, composition, and interoperation. However, the complexity of the OWL-S grammar makes it difficult to construct an OWL-S ontology manually. The model driven approach based on UML [3,4,5] has proven successful in modelling large and complex software systems.

In our proposed approach which aims to facilitate and raise the degree of automation in the semantic association process for the web services, we reverse engineer WSDL files into UML, annotate the generated models with relevant domain ontologies, and then we use a conversion tool to generate OWL-S descriptions from UML models. Our approach features the use of model driven architecture (MDA) concepts and strategies which enable the separation of the application design (or business logic) from the implementation platform. Specifically, these concepts allow us to translate XMI specifications [6] (e.g., XML encodings of UML) into OWL-S via eXtensible Stylesheet Language Transformations (XSLT) transformations [7]. We demonstrate our approach using CongoBuy example [2], and validate the correctness of our transformations using the OWL-S API [8].

The rest of the paper is organized as follows: In section 2, we describe the architecture and the main step of our approach. Section 3 introduces the UML profile for semantic web service; where section 4 details the reverse engineering process to generate UML model from WSDL files; section 5 shows the transformation rules between UML and OWL-S; Section 6 covers related work; and finally section 7 concludes the paper.



## 2 Our approach

This section describes our approach for using MDA techniques to synthesise OWL-S specification. The proposed methodology is broken up into three main steps (see Fig.1). The methodology will be explained briefly in this section and detailed description will be given in the subsequent sections.

### 2.1 Reverse engineering process

The developer imports the web service description into UML by a reverse engineering transformation. The developer can use a UML tool to review the imported models. Two UML diagrams will be generated from reverse engineering process; UML class diagram which describes the interface of service and UML activity diagrams of which the main part is obtained automatically when the order in which the activities will be performed is expressed by the developer by using data flow capabilities. The UML activity diagram is useful to describe the composite process of OWL-S process model.

### 2.2 Annotation process

The generated activity UML models constitute the skeleton UML profiles. The remainder of information will be enriched by as set of information from: UDDI, ontologies and the developer. Information in UDDI is useful to fill UML profiles part relating to owl-s profile (category name, taxonomy ...). In addition information like category helps the designer to identify appropriate domain ontologies to semantically annotate the UML models. Domain ontologies may be searched at organisations or business interest communities. For example the North American industry classification system represents such an effort and provides an ontology of products and services.

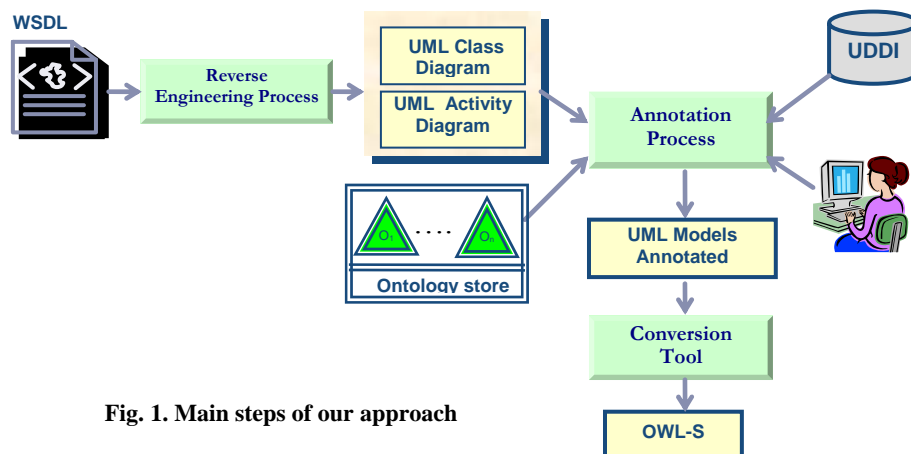


Fig. 1. Main steps of our approach

After the modeller has determined one or more candidate ontologies, he imports these into the UML models. Such a technique has already been proposed by work of Djuric to bring OWL ontologies into UML modelling environment [9]. Ontology is visualized with an interface UML as well as diagrams UML, the XSD complex types presented as classes in UML are mapped on ontology concepts, the parameters, the operations, the pre-conditions, the post-conditions and the effects are also linked and annotated. The category is filled starting from the UDDI. Once UML profile for semantic services Web is completed with a little additional effort on behalf of the developer, the third stage of transformation to OWL-S descriptions could be launched.

### 2.3 Conversion tool

Conversion tool implements a forward transformation from UML models obtained in step 2 to OWL-S. The transformation process is based on the eXtensible Stylesheet Language Transformations (XSLT) which is used in this project to convert an XML representation of an UML model into OWL-S description. The transformation rules in the Stylesheet rely on the UML profile we have created for OWL-S. The annotated UML models are used to generate a semantic web service description. This

description can be automatically generated by the transformation which we have explained further in section 5. The UML diagrams are exported in XMI format. The conversion tool is invoked to run multiple transformations on the input file to produce the corresponding Service, ServiceProfile, ServiceModel and ServiceGrounding OWL-S documents. The transformations come in the form of XSLT transformations which automatically convert XMI input into OWL-S specifications. From the standpoint of MDA terminology, the process of creating a UML diagram from WSDL description can be considered equivalent to reversely engineering a PSM to PIM. The XSLT transformations correspond to “Other Information” necessary to generate a second-level PIM in the form of an OWL-S specification.

### 3 UML profile for semantic Web Services

The UML profile includes stereotypes, custom data and tagged values used to extend the capabilities of the UML modelling language. A meta- model of UML profile is defined in [10] as an extension to UML 2.0 activity models by extending the elements shown inside the standard UML 2.0 package. UML ontology profile (UOP) [9,10] captures ontology concepts with properties and relationships. The UOP package contains the two elements (Ontology and OntClass). An OntClass element extends a UML class and represents a semantic concept. The most central concept in the profile is WebService which extends the UML Activity element. The other extensions of UML 2.0 activity models are introduced to support semantic annotation of web service elements. Table 1 contains a summary of all the new elements in UML profile.

*Table 1. Summary Of the UML Profile*

Stereotype	Extending UML meta model element	Tagged values	Usage
Web service	Activity	WsdL, Service, Port, Operation, URI_WSDL_document, URI version WSDL, URI WSDL port, URI WSDL service	Used to model a single web service operation. Its tagged values are sufficient to identify a web service operation and his binding.
Input	Activity inputpin	Name Input URI WSDL Input message	The stereotype is added to visualize if a pin is an input or an output
Output	Activity outputpin	Name Output URI WSDL Output message	Same as for <input>
Category	Comment	Taxonomy, TaxonomyURI, Value, Code	This item links the web service to a category defined by a Semantically defined concept within an ontology
Pre-condition	Constraint		A precondition is of constraint type and is visualised by a note in the diagram.
Post-condition	Constraint		Same as pre condition with respect to output parameters
Effect	Constraint		An effect statement defines the result of executing the web service
Binding operation	Comment	Binding name, style, type, transport	Used to identify binding operation

### 4 From WSDL to UML

In our approach, we have defined and implemented conversion rules from WSDL to UML to automate the reverse engineering process. We will investigate how to produce UML from WSDL by looking at an example web service with WSDL representation. The well-known “ExpressCongoBuy” provides a web service that allows customer to buy books. This section identifies two diagrams to modelling the WSDL content within UML: it’s expresses a UML class diagram which describes the interface of web service and its operations, and UML activity diagrams which specifies the internal behaviour of web service operation. We introduce bindings and access URIs in UML activity diagram useful in the transformation process from UML to OWL-S grounding. UML activity diagram is modelled according to UML profile presented in the previous section.

#### 4.1 UML class diagram

Fig. 3 shows a WSDL-UML class diagram of “ExpressCongoBuy” service. The service is mapped to UML class with stereotype “Business service”. The service realizes one interface “ExpressCongoBuyPortType”. This interface has one operation which takes parameters as input and gives a boolean response indicate if the book is successfully purchased and shipped or not. These parameters

refer to the two classes “CreditCard” and “SignInData”. We briefly explain the main conversion rules for transforming a WSDL file into WSDL-UML class diagram.

- A WSDL *service* is converted to a *BusinessService*.
- Each WSDL *port* within *service* is converted to realized relationship of the *BusinessService*.
- Each *portType* is converted to a UML interface.
- Finally, all XML schema defined types in the WSDL file are converted to UML classes.

## 4.2 UML activity diagram

UML activity diagram specifies the internal behaviour of web service operation. In our approach each operation in port type is transformed to an activity. The built-in control and data flow capabilities allows developer to define the order in which the activities will be performed. The UML activity diagram is obtained from WSDL file by applying the following rules:

- Each operation in port type is converted to single web service operation stereotyped “WebService” with eight tagged values (WSDL, Service, Port, Operation, URI\_WSDL\_document, URI version WSDL, URI WSDL port, URI WSDL service) which uniquely identify a web service operation within a WSDL file.
- The parts messages of input and output messages for operation are transformed to input and output parameters respectively stereotyped “Input” and “Output”.
- The Access URIs for input and output message are modelled with tagged values associated to stereotyped “Input” and “Output”.
- The binding operation information is presented as stereotyped note attached to WebService operation which extends the UML Comment element with four tagged values (BindingName, transport-protocols, style and type).

The UML activity diagram obtained after reverse engineering process represents a part of the diagram presented in Fig. 4. This diagram consists of a WebServiceOperation, the inputs and outputs parameters, stereotyped note for binding operation and three classes stemming from UML class diagram. These classes are mapped on existing OWL concepts during the annotation process of UML diagram. The rest of elements appearing in the diagram like category, pre-conditions, post-conditions and effects are defined and added in the annotation process.

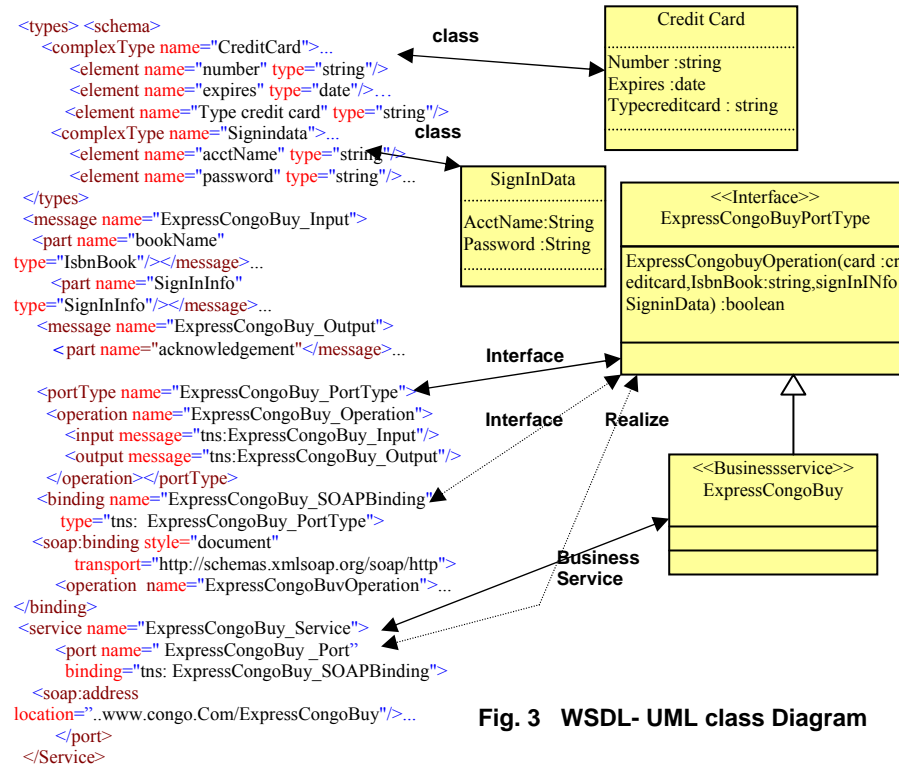


Fig. 3 WSDL- UML class Diagram

#### 4.3 Expresscongo buy example expressed in the UML profile

The model instance in Fig. 4 is used to explain our proposed UML profile by showing the OWL-S reference example ExpressCongoBuy in UML. In the ExpressCongoBuy example, there are five input parameters to identify the customer information (SignInInfo), Credit Card(creditCardNumber, creditCardType and creditCardExpirationDate) and the book (bookISBN). There are two mutually exclusive output parameters in the example. The first output parameter indicates that the book is successfully purchased and shipped to the buyer's address, while the second output provides a message informing that the book was out of stock. The parameter types are linked to syntactic and semantic types. In the ExpressCongoBuy example the parameters creditCardNumber and creditCardExpirationDate are syntactically defined by referring to standard XML Schema data types. The other parameters are defined with semantic types as UOP OntClasses and grouped inside UOP Ontology packages. A stereotyped note is attached to web service indicating the binding operation with tagged values (binding name, style and transport), these information are used in the transformation process for generating grounding.

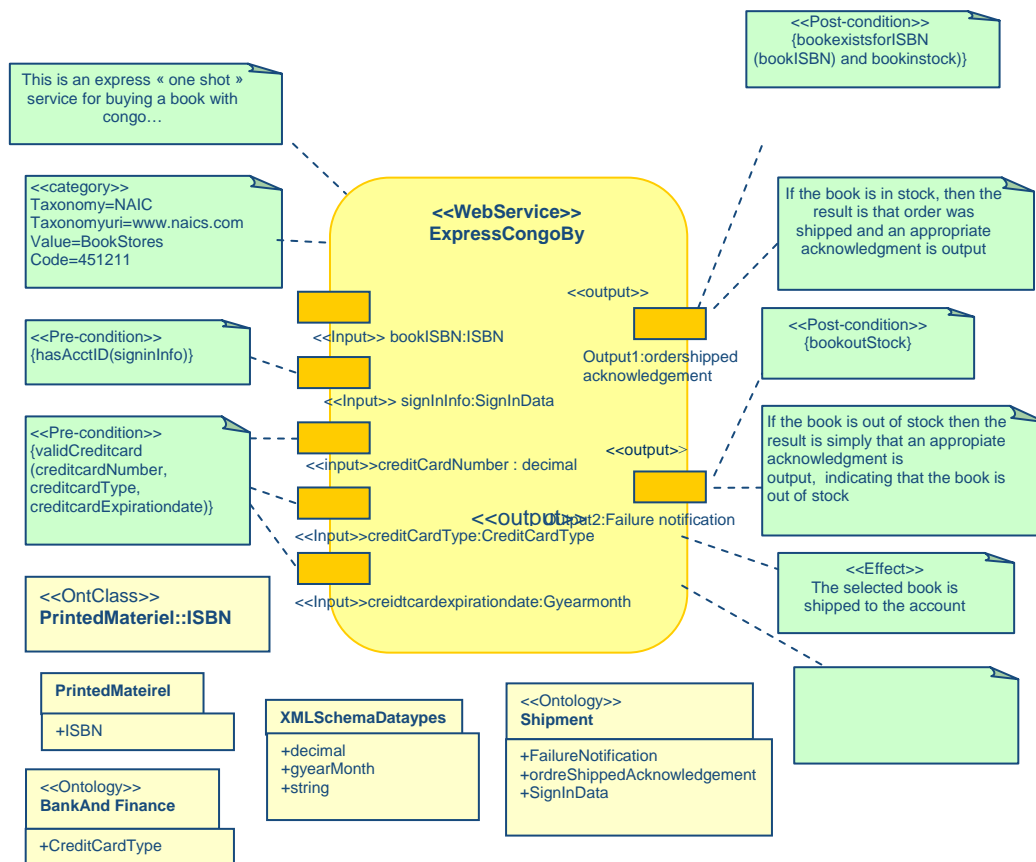


Fig. 4 ExpressCongobuy service represented in our UML profile

## 5 Transformations between our UML profile and OWL-S

This section describes the way UML models are transformed into an OWL-S ontology. An XMI file extracted from UML diagrams is transformed into owl-s ontology via an XSLT script. Table 2 is the map table for transforming XMI constructs into OWL-S ones. The leftmost column in the Table shows UML constructs, which are used to specify semantics. The middle column provides the XMI constructs for mapping UML constructs to OWL-S versions. Finally, the rightmost column shows the corresponding target constructs in an OWL-S ontology.

Inputs, outputs, pre- and post-conditions, and effects are generated at two places in the OWL-S document: Process section and Profile section. The transformation generates the elements in the Process part first. Then, these elements are basically duplicated for the Profile part. Reused ontologies are modelled in UML as separate packages with a URI as tagged value to identify the ontology. All such ontologies result in an import statement in the produced OWL-S document. In our approach, the bindings and access URIs are transported from WSDL document into UML activity diagrams in the form of extensible elements stereotypes and tagged values). So the process of generating the grounding is automatic, and as result we obtain a complete grounding.

**Table 2. Transformations relationship BETWEEN XMI and OWL-S**

UML	XMI	OWL-S
Action	<UML2:Activity>	Process
Pin	<UML2:InputPin>	Input
Pin	<UML2:OutputPin>	Output
Pin to Pin	<UML2:ActivityEdge>	{In Out}put Binding
comment	<UML:Comment>	categoryName
	<UML:Stereotype name='categoryName'>	
Tagged Value	<UML:TaggedValue>	taxonomy
	<UML:Stereotype name='taxonomy'>	
comment	<UML:comment>	serviceName
	<UML:Stereotype name='serviceName'>	
comment	<UML:comment>	textDescription
	<UML:Stereotype name='textDescription'>	
Constraint	<UML:Constraint>	Precondition
	<UML:Stereotype name='preCondition'>	
Constraint	<UML:Constraint>	Result
	<UML:Stereotype name='Result'>	
Constraint	<UML:Constraint>	Effect
	<UML:Stereotype name='Effect'>	
Ontology package	<UML : ontology>	External ontology
	<UML:stereotype name="ontologyName">	
OntoClass	<UML :ontoClass>	Owl class
	<UML: stereotype name=OntoclassName>	

## 6 Related work

The OWL-S editor [12] is a good standalone tool but is proprietary in nature and requires that the user become familiar with its user interface. Our approach leverages existing skills in UML modelling, which can greatly improve the efficiency of the semantic web service development workflow. Paolucci et al. developed the WSDL2OWLS system for automatically generating a one-shot OWL-S specification for a given WSDL file in [13]. As such their approach generates a complete grounding and incomplete profiles and process models. Jaeger, Engel and Geihs (2005) [14] proposed a methodology for developing semantic descriptions of Web services using OWL-S. They recognize the lack of tool support for the development of semantic descriptions. Our approach differs in that we use a hybrid approach, which combine between reverse engineering WSDL files into UML models and develop high-level OWL-S specifications in the context of an MDA environment.

## 7 Conclusion and futures investigations

Our approach facilitates the specification of semantic web services using model driven development. By importing WSDL descriptions of existing web services into UML diagrams, we show that UML can be used as a common integration platform. Our UML profile is expressive enough to capture and generate the needed semantic information of OWL-S and also can be reused towards different semantic web languages such WSML or WSDL-S [16] which extends WSDL2.0 with semantic descriptions, so OWL-S is just one candidate.

## 8 References

1. Chinnici, E.R, Moreau, J.J, Ryman, C.A, and Weerawarana, S. : Web service description language 1.1. W3C Note [Online] Available <http://www.w3.org/TR/2007/REC-wsdl20-20070626/>
2. Martin, D., and al. OWL Services Coalition. OWL-S: Semantic markup for web services, [Online] Available <http://www.daml.org/services/owl-s/1.0/owl-s.pdf> (2003).
3. OMG, 2002, Object Management Group's Model Driven Architecture [Online] Available [www.omg.org/mda](http://www.omg.org/mda).
4. Miller, J.M. et al. MDA guide version 1.0.1 (Tech. Rep. omg/2003-06-01). Object Management Group (2003).
5. Object Management Group (OMG). UML 2.0 Superstructure Specification, OMG Adopted Specification ptc/03-08-02. Technical Report (2003).
6. Object Management Group (OMG): XML metadata interchange (Tech. Rep. Formal/2005-09-01), OMG (2005).
7. Clark, J. : XSLT Transformations v1.0. W3C Recommendation [Online] Available <http://www.w3c.org/TR/xslt> (1999).
8. The MINDSWAP Group . OWL-S API. [online] available <http://www.mindswap.org/2004/owl-s/api>
9. Djuric D. : MDA-based Ontology Infrastructure. Computer Science Information Systems (ComSIS), (1(1):91–116 (2004).
10. Gronmo, R., Jaeger, M.C., and Hoff, H.: Transformations between UML and OWL-S, Springer-Verlag. Presented at The European Conference on Model Driven Architecture -Foundations and Applications (ECMDA-FA), Nuremberg, Germany (2005).
11. Horrocks, I., and al.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Technical Report, <http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/>, (2004).
12. Scicluna, J., Abela, C., and Montebello, M., : Visual modeling of owl-s service, Proceedings of the IADIS International Conference WWW/Internet, (2004).
13. Paolucci, M., Srinivasan, N., Sycara, K., & Nishimura ,T. : Towards a semantic choreography of Web services: From WSDL to DAML-S, Proceedings of the International Conference on Web Services. IEEE (2003.)
14. Jaeger, M.C, Engel, L., and Geihs, K.: A methodology for developing owl-s descriptions, Proceedings of the First International Conference on Interoperability of Enterprise Software and Applications Workshop on Web Services and Interoperability (2005).
15. Rajasekaran, P., Miller, J.A., Verma, K., and Sheth, P.A. : Enhancing Web Services description and Discovery to Facilitate Composition. In Semantic Web Services and Web Process Composition, Proceedings of the First International Workshop, volume 3387 of Lecture Notes in Computer Science, San Diego, California, USA (2004).

## Vers une plateforme de modélisation agile et d'analyse des processus métier

Mohamed Boukhebouze, Youssef Amghar, Aïcha-Nabila Benharkat

LIRIS, Institut National des Sciences Appliquées de Lyon, France  
{mohamed.boukhebouze, youssef.amghar, nabila.benharkat}@insa-lyon.fr

**Résumé:** Dans ce papier nous allons parler de la plateforme BP-FAMA (Business Process Framework for Agility of Modelling and Analyses) qui est née, d'une part du besoin d'avoir un niveau abstraction ou les préoccupations algorithmiques d'un côté et les règles de gestion, de l'autre, sont souvent mélangées ce qui rend difficile l'évolution des processus, et d'autre part du besoin d'analyser la qualité du passage d'une spécification de processus de haut niveau à un processus exécutable. Ces objectifs ne vont pas sans l'exigence de nouveaux langages pour une modélisation agile : BPAMN<sup>1</sup> et un nouveau langage d'exécution agile BPAEL<sup>2</sup>.

**Mots clés:** la modélisation des processus métier, l'agilité de la modélisation, les règles métier.

### 1 Introduction

La modélisation des processus constitue donc une étape importante dans la gestion des processus métiers, car elle permet de spécifier les connaissances du métier d'une entreprise. Pour cette raison, ces entreprises doivent se baser sur des langages puissants afin de donner une description complète des différentes dimensions d'un processus métier. Dans ce contexte deux standards sont proposés : une notation graphique commune à tous les outils de modélisation appelée BPMN, et un langage d'exécution du processus appelé BPEL et cela afin de garantir une portabilité des processus réalisés sur différentes plateformes.

Ces spécifications BPMN et BPEL sont maintenant stables et se sont adaptées aux besoins du marché. Plusieurs éditeurs les ont adoptées et introduites dans leurs outils. Toutefois, ces spécifications nécessitent d'être vérifiées avant leur implantation, malheureusement, ces standards se focalisent davantage sur le niveau descriptif du métier, notamment des aspects fonctionnels d'un processus, sans offrir des mécanismes de support à la vérification des spécifications. En effet, la fiabilité des processus d'une part et la maîtrise des coûts de maintenance d'autre part conduisent à accorder une attention plus grande aux questions de vérification des processus.

En outre, les règles métier sont des connaissances collectives impliquant des représentations partagées concernant les comportements ou les modèles d'activité (ce sont les règles organisationnelles), les représentations communes sur l'environnement par exemple (ce sont les faits), et le langage utilisé pour pouvoir communiquer et interpréter les règles et les faits. Par la nature dynamique de l'environnement, les faits changent; les règles changent car l'organisation peut changer et aussi le langage car il correspond à des interprétations successives faites dans le temps.

Nous situons notre problématique à deux niveaux : (1) Niveau abstraction où les préoccupations algorithmiques d'un côté et les règles de gestion, de l'autre, sont souvent mélangées ce qui rend difficile l'évolution des applications : l'agilité. (2) Le niveau Analyse : même s'il existe des outils permettant le passage d'une spécification de processus de haut niveau à un processus BPEL, rien ne garantit la qualité du processus généré.

Dans ce papier nous allons décrire l'architecture de la plateforme BP-FAMA (Business Process Framework for Agility of Modelling and Analyses) qui fait l'objet d'un développement au sein de notre laboratoire et qui permet de répondre à ces deux problématiques. Pour cela nous allons introduire dans la section 2 les motivations qui nous ont poussées à proposer cette plateforme. Dans la section 3 nous

---

<sup>1</sup> BPAMN : **B**usiness **P**rocess **A**gile **M**odeling **N**otation

<sup>2</sup> BPAEL : **B**usiness **P**rocess **A**gile **E**xecution **L**anguage

détaillons l'architecture de cette plateforme, la section 4 est consacré à l'explication des deux langages proposés BPAMN et BPAEL pour BP-FAMA les deux dernières sections décrivent un cas d'étude illustratif et une partie discussion.

## 2 Motivations

Pour atteindre notre objectif, nous nous sommes intéressés à répondre à deux questions : comment modéliser un processus métier d'une manière agile et comment peut-on analyser un processus métier ?

Premièrement, la modélisation des processus constitue une étape importante dans le cycle de gestion d'un processus métier car elle permet la spécification des connaissances métier d'une entreprise. Pour cette raison, des langages puissants sont nécessaires pour donner une description complète des différents aspects d'un processus métier. Dans ce contexte, plusieurs langages graphiques qui permettent de décrire des processus ont été proposés tels que UML [1], YAWL [2], BPMN [3]. Des langages d'exécution des processus basés sur XML ont été aussi proposés tels que XPDL [4], et BPEL4WS [5]. En parallèle, plusieurs travaux ont été menés pour comparer ces langages d'exécution tels que [6]. Ces études démontrent que BPEL est avantageux parce qu'il permet de représenter un plus grand nombre d'éléments de base des processus métier par rapport à ses prédécesseurs. Pour cette raison, BPEL est devenu le langage le plus utilisé dans industrie.

Toutefois, en utilisant ces langages graphiques et d'exécution, les concepteurs implémentent les règles métier dans le code du processus métier, cette pratique rend les processus métier rigides et difficiles à maintenir. En effet, la nature dynamique des environnements des organisations, rend les règles susceptibles d'être modifiées ou supprimées. Donc l'intégration de la logique algorithmique et les règles métier dans le même code des processus compromet l'agilité de la modélisation. Pour remédier à ce problème, le repérage et l'identification des règles métier au sein d'une modélisation d'un processus métier nous semble nécessaire et importante pour permettre une évolution et une maintenance des règles de façon indépendante du processus et permettre ainsi le partage de ces règles par d'autres processus.

Par ailleurs, les entreprises doivent se baser sur des processus métier robustes et fiables pour mener à bien leurs missions. La fiabilité des processus est une question primordiale car ces derniers automatisent tout ou partie de la chaîne de valeur d'une entreprise et capitalise en même temps sur leur système d'information. Un processus erroné peut avoir, pour une entreprise, des conséquences graves sur le plan économique. Pour cela, la vérification par modèles formels est considérée comme la technique la plus utilisée pour repérer les erreurs de spécifications lors d'une conception d'un processus métier. D'une façon générale, le réseau de Pétri (RdP) est le modèle le plus utilisé [7, 8]. Ceci étant, d'autres approches, se basant sur des modèles différents, ont été proposées. Le travail [9] de Breugel recense les différents travaux proposés pour la vérification des processus BPEL

Cependant, le changement d'un élément du processus conduit à une re-vérification du processus dans ça globalité. D'autre part, ces techniques ne se focalisent que sur la cohérence du système et ne garantissent pas que le processus implémenté réponde réellement aux attentes des concepteurs.

Dans ce qui suit nous allons parler de notre plateforme BP-FAMA qui contribue à résoudre aux problèmes précédant.

## 3 La plateforme BP- FAMA

L'objectif général de la plateforme FAMA est d'améliorer le fonctionnement des BPMS (voir figure1). En effet, cette plateforme permet de favoriser l'agilité de la modélisation du processus métier, favoriser l'agilité de l'exécution du processus et intégrer une étape de vérification et cela pour chaque phase de cycle de vie d'un processus métier.



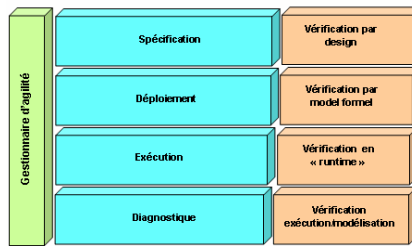


Fig1- L'Architecture globale de la plateforme BP-FAMA.

### 3.1 La phase de spécification

Dans la phase de spécification, le concepteur définit les éléments qui constituent le processus métier où il redéfinit les éléments d'un processus existant dans le but de l'améliorer. Cette définition consiste en un moyen de dialogue entre les responsables des processus et les équipes opérationnelles en charge de les exécuter. Pour cela, des langages graphiques sont utilisés dans cette phase pour décrire les modèles de processus parce qu'ils offrent un moyen convivial de représentation et facilite la compréhension du processus. Dans la plateforme FAMA (voir figure 2), le langage utilisé pour décrire graphiquement un processus métier est le nouveau langage BPAMN. Ce langage permet de représenter un processus métier avec une notation graphique complète (éléments graphiques et diagrammes) héritée de son ancêtre le standard BPMN.

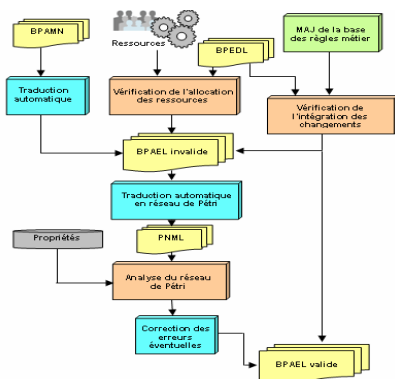


Fig2- Architecture de la spécification

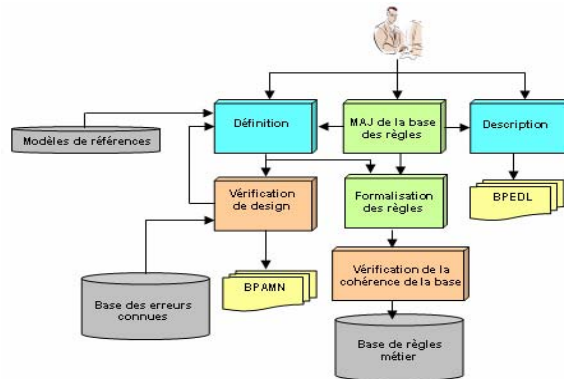


Fig3 Architecture de déploiement

Le grand atout de ce nouveau langage est qu'il permet de séparer la définition des règles métier par rapport à la logique algorithmique du processus métier. Cette séparation permet de gérer indépendamment les règles métier dans une base dédiée, et ainsi permettre le partage de ces règles par plusieurs processus. La vérification est effectuée à chaque mise à jour afin d'assurer la cohérence de cette base de règle. Pour cela, la formulation des règles métiers doit être rigoureuse, concise et précise pour garantir que ces règles soient non ambiguës, cohérentes, complètes et énoncées avec une terminologie commune au métier. En parallèle, une vérification anticipée est effectuée pour repérer les erreurs connues dans la modélisation du processus.

### 3.2 La phase de déploiement

Dans la phase déploiement, le processus modélisé est mis en œuvre en développant les applications métier qui l'implémentent, dans le contexte de l'organisation et de son système d'information, et en allouant les ressources nécessaires pour l'exécution de différentes activités du processus. De cette façon, le processus devient opérationnel. Dans la plateforme FAMA (voir figure 3), le processus est mis en œuvre à partir de la modélisation exprimé en BPAMN et cela en traduisant automatiquement, cette modélisation, en un nouveau langage d'exécution appelé BPAEL. En effet, BPAEL se base entièrement sur le standard BPEL auquel on ajoute une nouvelle activité structurée « Rule » qui nous aide à identifier

les règles métier dans un processus BPAEL. De cette façon, les mises à jour des règles peuvent être intégrées dans le processus automatiquement. En parallèle, le langage BPAEL, comme son ancêtre BPEL, n'offre pas des mécanismes de support à la vérification des spécifications des processus. Pour cela, nous avons eu recours aux modèles formels pour repérer les erreurs fonctionnelles éventuelles. Nous avons opté pour les réseaux de Pétri (RdP) et cela à cause de leur grande capacité à modéliser une grande variété de processus métier. En plus, les RdP offrent aussi un large éventail de propriétés mathématiques qui permettent d'analyser le bon fonctionnement du processus. Pour cela, un outil de réécriture de la spécification des processus en terme de RdP (des traducteurs) est proposé, les réseaux obtenus sont exprimés dans le langage PNML (Petri Net Markup Language) qui est utilisé par de larges outils d'analyse de RdP (analyseurs). Ces derniers permettent de vérifier certaines propriétés comme l'absence de blocage, la vivacité du réseau, etc., ils peuvent ainsi détecter les erreurs en tenant compte des propriétés détectées.

### 3.3 La phase d'exécution

Dans la phase d'exécution, le langage d'exécution des processus métier qui spécifie le déroulement de l'ensemble des activités d'un processus est interprété par des moteurs d'exécution en automatisant les interactions entre les participants du processus (les documents, les informations et les tâches) et en faisant appel aux différentes ressources. Dans le but d'intercepter les exceptions non modélisées par le concepteur (comme une panne matérielle ou une indisponibilité d'une ressource), la vérification est lancée, dans la plateforme FAMA, d'une manière parallèle avec l'exécution (voir figure 4). De cette façon, on essaye de réagir face à la situation qui viserait à déstabiliser le fonctionnement du processus métier et éviter de faire appel à l'utilisation de la vérification par le modèle RdP qui est coûteux en terme de temps (algorithme de traduction plus algorithme d'analyse). Finalement, les différents fichiers logs et les traces cumulées lors de l'exécution d'un grand nombre d'instances de processus sont stockées dans une base pour être utilisées dans la prochaine étape.

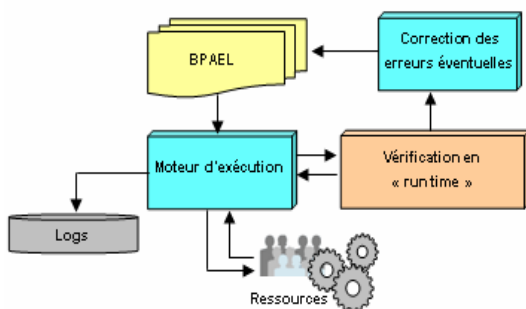


Fig4- L'architecture de l'exécution

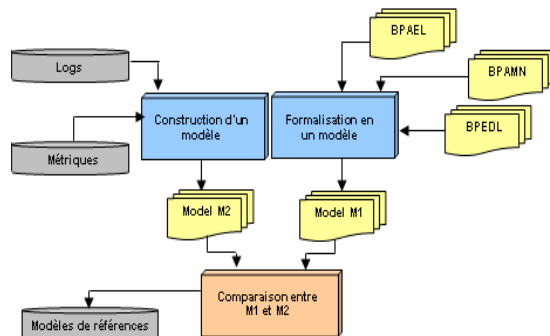


Fig5- L'architecture du diagnostic

### 3.4 La phase diagnostic

Dans la phase de diagnostic, le processus exécutable est analysé dans le but de mesurer les performances opérationnelles en se basant sur les fichiers logs. Dans la plateforme FAMA, le processus exécutable est analysé afin de reconstruire, à partir des informations cumulées lors de la phase précédente (voir figure 5). Le processus résultat sera comparé au processus modélisé pour savoir qui s'agit du même déroulement de l'ensemble des activités.

Dans ce papier nous allons détailler les deux nouveaux langages proposés : le langage graphique BPAMN et le langage d'exécution BPAEL.

## 4 Les nouveaux langages proposés

Le BPAMN vient pour ajouter un aspect agilité à une représentation BPMN. En effet, ce nouveau langage permet l'identification des règles métier dans une modélisation à l'aide de nouveaux symboles (des croix voir la partie 5) dédiés à modéliser les éléments des règles métier séparément des autres éléments des

processus métier. De cette manière, le repérage de ces règles est possible tout au long du cycle de vie d'un processus métier.

```
<Rule RuleName = "rname" >
  <onMessage partnerLink="rname" portType="qname"
    operation="rname" variable="rname" ?>*
    <Rule condition="bool-expr">+
      Activity+
    </Rule condition>
  </onMessage>
  <onAlarm (for="duration-expr" | until="deadline-expr")>*
    <Rule condition="bool-expr">+
      Activity+
    </Rule condition>
  </onAlarm>
</Rule>
```

Fig. 6 La déclaration de l'activité RULE de BPAEL

Le nouveau langage d'exécution BPAEL est proposé dans le but de tracer les règles métier dans le processus opérationnel. En effet, le BPAEL est basé entièrement sur la norme BPEL, auquel nous ajoutons une nouvelle activité structurée appelée *Rule* pour repérer l'emplacement des règles métier dans un processus BPAEL. Nous considérons une règle métier sous la forme : ON [Événement] IF [conditions] THEN [activité]. Pour cette raison, la nouvelle activité *Rule* combine le fonctionnement de l'activité *Pick* qui permet d'intercepter les événements et le fonctionnement de l'activité *If* qui permet d'exprimer des conditions. Un identificateur de règle est ajouté dans cette nouvelle activité afin de conserver une trace de la règle en question dans la base de règle (voir la figure 6).

## 5 Etude de cas

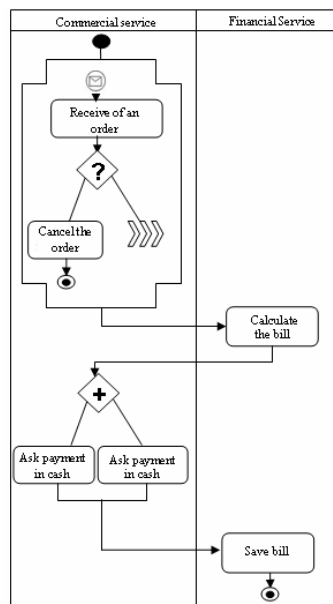


Fig7- Une représentation BPANM du processus de traitement de commande

```
<process name="PurchaseOrder">
  <sequence>
    <Rule IDName="R1">
      <onMessage partnerLink="ComercialService"
        <RuleCondition>
          Bpel: IsClientIsRegistered("ClientName") = false
        <RuleCondition>
          <Reply "Purchase order Cancelled"/>
        </onMessage>
      </Rule>
    <invoke operation="requestBillPayment"/>
    <if>
      <Condition>
        Bpel: PaymentInCash() = true
      <Condition>
        <invoke operation="Save Payment Bill in cash"/>
      </if>
    <else>
      <invoke operation="Save Payment Bill in credit card"/>
    </else>
    </sequence>
  </process>
```

Fig8- Un processus BPAEL de traitement de commande

Pour illustrer le nouveau symbole graphique de règle du langage BPAMN (figure 7) et la nouvelle activité *Rule* du langage BPAEL (figure 8), A la réception d'une commande d'un client, ce processus calcule le prix de la facture et envoie cette dernière au client, pour effectuer un paiement en espèces ou avec une carte bancaire. La facture est finalement enregistrée. Une règle métier est alors définie, elle concerne le client qui doit être enregistré dans la base de données pour pouvoir lancer une commande. La figure 7 représente une modélisation graphique, exprimée en BPAMN, du processus de traitement de commande.

La règle représentée par la croix de la figure se déclenche lors que la réception d'une commande (événement réception message est représenté par le symbole associé), si le client est enregistré (cette condition est exprimée par le symbole associé) alors on continue à exécuter le processus. Pour exprimer ce mécanisme on utilise le symbole **Link**. Dans le cas inverse (si le client n'est pas enregistré) le processus s'arrête en annulant la commande.

## 6 Discussion

Nous avons proposé dans cet article un Framework appelé BP-FAMA qui tente de répondre à ces deux problèmes. En effet nous pensons que les règles métier doivent être identifiées par le concepteur lors de la phase de spécification et la phase de déploiement du processus. Le manque de ce mécanisme d'identification de règles dans les deux standards BPMN et BPEL nous a poussé à proposer des extensions à ces standards : le BPAMN un langage graphique de modélisation agile des processus et BPAEL un langage d'exécution agile de processus. Cette identification de règle au sein d'un processus métier permet de garder trace de ces règles pour une gestion séparée de la logique algorithmique de processus et permet aussi d'intégrer les mises à jour automatiquement.

Nous pensons aussi que pour offrir une analyse complète à un processus métier, l'intégration d'une étape de vérification à chaque phase de cycle de vie d'un processus métier. En effet, dans la phase de spécification, le concepteur doit être assisté pour modéliser le processus et cela en détectant les éléments qui peuvent être une source potentielle d'erreurs. Dans la phase de déploiement, la cohérence fonctionnelle du processus modélisé est vérifiée par un modèle formel. Dans la phase d'exécution, la vérification est lancée en démon pour intercepter les exceptions éventuelles et essayer de réagir dans le but à conduire l'exécution du processus vers une situation stable ou valide. Dans la phase de diagnostic, le processus est reconstruit à partir des informations cumulées lors de l'exécution d'un grand nombre d'instances de processus dans le but de s'assurer que ce qu'on a modélisé correspond à ce qui est entrain d'être exécuté.

## 7 References

- 1- Object Management Group: Business. Unified Modeling Language: Superstructure version 2.1.1 formal/2007-02-03 (2003).
- 2- W.M.P. van der Aalst and A.H.M. ter Hofstede. YAWL: Yet Another Workflow Language. QUT Technical report, FIT-TR-2002-06, Queensland University of Technology, Brisbane, (2002)..
- 3- Object Management Group, Business Process Modeling Notation (BPMN) Specification Final Adopted Specification dtc/06-02-01 (2003).
- 4- The Workflow Management Coalition: Workflow Management Coalition Workflow Standard Process Definition Interface: XML Process Definition Language. In the specification rapport (2005).
- 5- OASIS: Business Process Execution Language for Web Services (BPEL4WS): Version 2.0. In BPEL4WS specification report (2007).
- 6- Vasko, M., Dustdar, S.: A view based analysis of workflow modeling languages. In the 14th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP'06) (2006).
- 7- Ouyang, C., Verbeek, E., van der Aalst, W.M.P., Breutel, S., Dumas, M. ter Hofstede, A.H.M., 2005. WofBPEL: a tool for automated analysis of BPEL processes. In ICSOC 2005, International Conference of Service-Oriented Computing, Berlin.
- 8- Yang, Y., Tan, Q. Yu, J., Liu, F. 2005. Transformation BPEL to CP-nets for verifying web services composition. In the International Conference on Next Generation Web Services Practices, Korea.
- 9- Van Dongen, B.F., Jansen-Vullers, M.H., 2005. Verification of SAP Reference Models. In BPM 2005, 3rd International Conference, France.

## Agents mobiles et web services pour le E-Commerce

Linda Zeghache<sup>1</sup>, Aouaouche El-Maouhab<sup>1</sup>, Sanaa Medjedba<sup>1</sup> and Habiba Boulefaat<sup>1</sup>,

<sup>1</sup> Centre de Recherche sur l'Information Scientifique et Technique (CERIST), Division Technologies des Réseaux d'Information (DTRI)

03 rue frères Aissiou, Benaknoun, Alger

l.zeghache@dtri.cerist.dz, elmaouhab@wissal.dz, sanaamedjadba@yahoo.fr, boulefaat-habiba@hotmail.com

**Résumé.** Les agents mobiles sont des programmes autonomes qui peuvent se déplacer à travers un réseau hétérogène sous leur propre contrôle. Ils sont caractérisés par leur dynamique, évolutivité et autonomie. D'autre part, les services Web se présentent aujourd'hui comme un support crédible permettant à des applications d'exposer leurs fonctionnalités au travers d'interfaces standardisées et de plus en plus éprouvées. La particularité des services Web réside dans le fait qu'elle utilise la technologie Internet comme infrastructure de communication. On assiste aujourd'hui à un rapprochement entre les recherches en agents mobiles et les services Web. Dans le présent travail nous proposons d'utiliser des services Web comme un cadre architectural et/ou technologique pour mettre en place des agents mobiles accessibles à travers le web pour une application de commerce électronique.

**Mots-Clés:** Agent Mobile, Système d'agents mobiles, Web services, Commerce électronique

### 1 Introduction

Récemment, la recherche en systèmes répartis a vu l'émergence d'un nouveau modèle pour la structuration d'applications réparties : La programmation par agents mobiles. Les agents mobiles visent principalement des applications réparties sur des réseaux à grande distance, car ils permettent de déplacer l'exécution vers les serveurs et de diminuer ainsi le coût d'accès à ces serveurs.

D'autre part, le développement des systèmes distribués s'est vu passer d'architecture orientées systèmes à des Architectures Orientées Services (AOS) facilitant l'interopération et l'intégration de composants logiciels. Les services Web sont aujourd'hui la principale implémentation des AOS [4].

Les services Web se présentent aujourd'hui comme un support crédible permettant à des applications d'exposer leurs fonctionnalités à travers des interfaces standardisées et de plus en plus éprouvées. La particularité des services Web réside dans le fait qu'ils utilisent la technologie Internet comme infrastructure pour la communication entre les composants logiciels (les services Web) et ceci en mettant en place un cadre de travail basé sur un ensemble de standards.

Les services Web reposent sur l'ubiquité de l'infrastructure d'Internet. L'interopérabilité est donc une caractéristique intrinsèque aux services Web. D'un autre côté, les agents mobiles sont caractérisés par leur dynamique, évolutivité et autonomie contrairement aux services web qui sont statiques.

### 2 L'approche Agents Mobiles

Le concept d'agent mobile est un peu particulier. Il repose non plus sur le modèle classique client-serveur mais sur un modèle d'exécution distribué.

Un agent mobile est un agent qui peut se déplacer à travers un réseau hétérogène sous son propre contrôle, migrant d'un site vers un autre et interagissant avec d'autres agents et ressources sur chacun d'eux, puis éventuellement revenir à son site natal une fois sa tâche accomplie [1].

Les agents mobiles offrent plusieurs avantages améliorant la performance de plusieurs applications distribuées. Cette amélioration peut être ressentie dans la réduction du trafic réseau, la répartition dynamique de charge, la commodité par rapport aux programmeurs ou simplement dans l'habilité de continuer l'interaction avec un utilisateur durant une déconnexion du réseau [1].

Les domaines d'application des agents mobiles sont multiples. Tout ce qui est fait en mode client/serveur pourra être fait avec des agents mobiles. Combien même une partie des applications ne

nécessitent pas ce nouveau modèle, certaines se révéleront plus efficaces en l'utilisant tel que le e-commerce, la recherche d'information, le calcul parallèle et les télécommunications [2][3].

L'utilisation des agents mobiles présente plusieurs avantages. Néanmoins, ils posent quelques problèmes, dont les plus importants sont ceux relatifs à l'interopérabilité, la sécurité et la tolérance aux pannes des systèmes d'agents mobiles.

### 3 Les Web Services

Un Service Web est une application accessible à partir du Web. Il utilise les protocoles Internet pour communiquer et utilise un langage standard pour décrire son interface [5].

Les Web Services sont les principaux blocs de construction de l'informatique distribuée sur Internet. Les normes ouvertes et l'accent mis sur la communication et la collaboration entre les gens et les applications ont créé un environnement dans lequel les Web Services deviennent la plate-forme pour l'intégration d'applications. Les applications sont construites à l'aide de plusieurs Web Services provenant de sources différentes et fonctionnant ensemble, quel que soit leur emplacement et la manière dont elles ont été mises en œuvre [4].

Les Services Web prennent leur origine dans l'informatique distribuée et dans l'avènement du Web. Le but de l'informatique distribuée est de permettre à une application sur une machine d'accéder à une fonction d'une autre application sur une machine distante et ce, de la même manière que l'appel d'une fonction locale, indépendamment des plates-formes et des langages utilisés.

Le point qui a favorisé l'essor des Services Web, c'est le développement d'Internet par la démocratisation du haut débit, la structuration des données via XML (Extensible Markup Language) et la recherche d'interopérabilité [4].

### 4 Rapprochement Agents et Web Services

On assiste aujourd'hui à un rapprochement entre les recherches sur les agents et les Services Web [6]. Ce rapprochement se manifeste sous différents aspects. On en distingue essentiellement deux : d'une part l'utilisation des agents en tant que cadre conceptuel pour mettre en place des Services Web sophistiqués, ou encore des outils de planification et composition de Services Web. D'autre part l'utilisation des Services Web comme un cadre architectural et/ou technologique pour mettre en place des agents accessibles à travers Internet.

Concernant le premier aspect, des travaux dans [7] proposent une implémentation des Services Web suivant le modèle agent afin de faciliter la gestion des transactions entre Services Web commerciaux.

Quant à l'utilisation des Services Web comme environnement technologique et architectural pour les agents, les travaux ne sont pas aussi présents que dans le contexte précédent. Les auteurs dans [6][4] présentent une éventualité du rapprochement entre les services Web et les agent FIPA<sup>3</sup>. Ils définissent la notion d'agent Services Web. Les agents Services Web sont des agents FIPA qui publient leurs capacités sous forme de services Web WSDL en procédant à une traduction de leur description en vue d'être invoqués par d'autres applications non agent. Cependant la plupart des travaux proposés concernent essentiellement les systèmes multi-agents basés sur la norme FIPA.

L'objectif principal de ce travail est de rendre les capacités des agents mobiles accessibles à travers des Services Web pour une application de commerce électronique. L'application consiste à réaliser un robot magasinier (shopping robot) dont le rôle consiste à parcourir le Web, à la place du cyberconsommateur, à la recherche de sites commerciaux offrant le produit désiré au meilleur prix.

### 5 Architecture Proposée

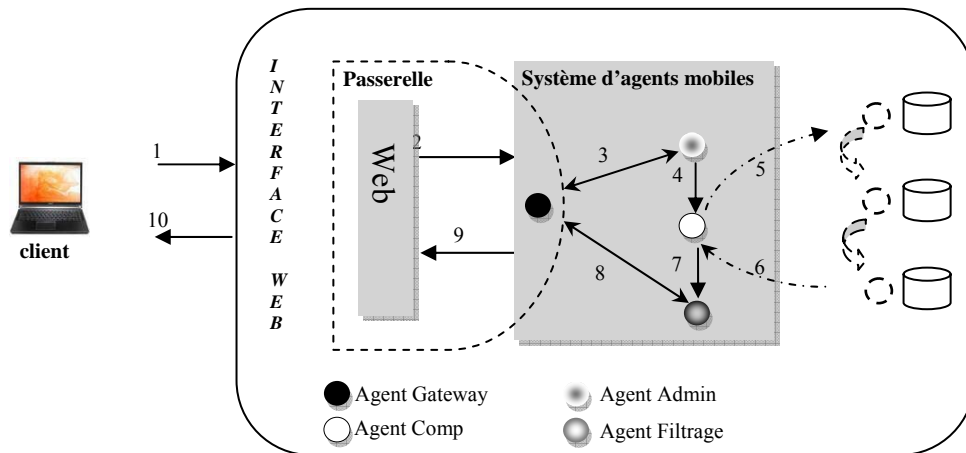
Nous avons développé un robot magasinier basé sur des agents mobiles qui parcourent plusieurs sites marchands à la recherche d'un produit donné et retournent le résultat au client sous forme d'un tableau

---

<sup>3</sup> Foundation for Intelligent Physical Agents, norme d'interopérabilité des systèmes multi-agents.

comparatif contenant les différentes caractéristiques du produit ainsi que les liens vers les différents sites marchands qui vendent le produit.

L'architecture proposée est structurée en trois niveaux (Fig1) :



**Fig. 1.** Schéma de l'architecture proposée. 1 : envoi de la requête client via une interface web. 2 : lancement de l'agent Gateway. 3 : l'agent Gateway envoie la requête à l'agent Admin. 4 : l'agent Admin crée les agents Comp. 5 : l'agent Comp migre vers les sites marchands pour chercher les caractéristiques du produit demandé par le client. 6 : retour de l'agent Comp. 7 : envoi du résultat à l'agent Filtrage. 8 : l'agent Filtrage tri les résultats et envoie la réponse à l'agent Gateway. 9 : l'agent Gateway envoie le résultat de la requête au web service. 10 : affichage du résultat au client.

**Niveau 1 :** représente le côté client, qui dispose d'une interface web lui permettant d'introduire sa requête.

**Niveau 2 :** représente la passerelle qui permet de transférer la requête client aux agents mobiles. Elle est composée d'un web service et d'un agent Gateway.

- Le web service permet de publier les capacités de l'agent mobile et de recevoir les requêtes destinées à cet agent.

- L'agent Gateway : joue le rôle d'un relais entre les agents et le web service.

Le client perçoit l'agent comme un web service, il l'invoque en utilisant les standards du web. Quand le web service reçoit la requête il crée un agent Gateway qui se charge de transmettre la requête client à l'agent mobile.

L'agent Gateway reste en attente de la réponse et envoie le résultat au web service pour l'afficher au client.

**Niveau 3 :** composé d'un ensemble d'agents mobiles qui coopèrent entre eux pour satisfaire la requête du client. Trois types d'agents sont définis :

- L'agent Admin, il gère les différents agents se trouvant sur le système d'agents.
- L'agent Comp, il migre vers les différents sites marchands pour chercher l'information requise.
- L'agent Filtrage : il se charge du filtrage et de la fusion des résultats retournés par les agents Comps.

## 6 Implémentation

Nous avons choisi d'implémenter notre application avec le langage JAVA vu que ce dernier est un langage simple, intuitif, orienté objet, performant et dynamique.

Dans le même contexte nous avons opté pour JADE (Java Agent Development Framework) [8] dans notre choix de la plate forme d'agents. Ce choix est justifié par le fait que JADE est une plate-forme multi-agents développée en Java, dotée d'une interface graphique et peut être répartie sur plusieurs serveurs.

Pour la plateforme web services, nous avons utilisé l'architecture J2EE [9] qui est basée sur le langage Java qui permet le déploiement de composants sur diverses plates-formes indépendamment de leur langage de programmation.

## 7 Evaluation

Cette architecture permet de publier les services offerts par les agents à travers le web afin de faciliter leurs invocations. Ces agents, qui nécessitent une infrastructure d'exécution et des langages spécifiques de communication, peuvent grâce à cette passerelle interagir avec des applications non-agents dans un environnement web.

Des travaux similaires ont été effectués dans [8] où la plateforme JADE implémente une passerelle qui effectue la transformation des messages SOAP vers des messages FIPA-ACL compris par les agents et vice-versa afin d'assurer l'échange de données entre les applications et les agents à travers le web. Cependant, cette solution reste spécifique aux agents conformément au standard FIPA.

Notre solution permet d'effectuer le transfert des requêtes clients vers les agents à travers les web services indépendamment des langages de communication d'agents.

## 8 Conclusion

Les agents mobiles semblent ouvrir une nouvelle voie pour la conception future d'applications réparties grâce aux multiples avantages qu'ils offrent notamment leurs qualités asynchrones et leur utilisation modérée des ressources du réseau.

Les agents mobiles ont besoin de leur propre infrastructure et plateforme pour s'exécuter et interagir avec les autres applications. Ces plateformes sont hétérogènes et posent un problème d'interopérabilité.

Le développement d'une passerelle web services permet aux applications d'accéder aux agents mobiles et d'invoquer leurs services via le web. Cela assure une interopérabilité fonctionnelle entre les applications web et les agents mobiles.

Dans des travaux futurs nous proposons d'utiliser des passerelles web services afin d'assurer l'interopérabilité entre les systèmes d'agents mobiles hétérogènes.

## References

1. K. Rothermel, M. Schwehm, Mobile Agents, In: A. Kent and J. G. Williams (Eds.): Encyclopedia for Computer Science and Technology, New York: M. Dekker Inc., 1998.
2. Nicolas Esposito, Les agents mobiles – Une introduction, DASSAULT SYSTÈMES – Recherche et nouvelles technologies, France, 2000
3. T. Gherbi, Structuration d'applications réparties dans un environnement mobile, Thèse Magister, l'U.S.T.H.B., 2000.
4. T. Melliti. Interopérabilité des services Web complexes. Application aux systèmes multi-agents. Thèse de Doctorat de l'Université Paris IX Dauphine. 2004.
5. F. Curbera, W. A. Nagy, and S. Weerawarana. Web services: Why and how? OOPSLA 2001 Workshop on Object-Oriented Web Services, 2001.
6. M. Lyell, L. Rosen, M. Casagni-Simkins, and D. Norris. On software agents and web services: Usage and design concepts and issues. In Workshop on Web services And Agent-based engineering, Melbourne, Australia, 2003.
7. T. Jin and S. Goschnick. Utilizing model (abt). In Workshop on Melbourne, Australia, 2003.
8. Jade - Java Agent DEvelopment Framework. <http://jade.tilab.com/>
9. J2EE, Java 2 Enterprise Edition : <http://ashita-studio.com/articles/j2ee/>



# Heavy-tailed distribution and QoS in distributed Web server systems

Fouzi Semchedine, Louiza Bouallouche-Medjkoune, Djamil Aïssani  
 Doctoral School in Computer Science

and

LAMOS, Laboratory of Modelisation and Optimization of Systems  
 University of Béjaïa 06000, Algeria  
 fouzi\_jams@yahoo.fr, louiza\_medjkoune@yahoo.fr, djamil.aissani@lamos.org

**Abstract.** The Quality of Service (QoS) in distributed Web server systems focuses on the choice of the assignment policy of tasks that reach these systems in order to minimize their response time and thus, improve the performances. These tasks, generally, follow the heavy-tailed distribution which have the property that there is a tiny fraction (about 3%) of large tasks that makes half (50 %) of the total load. In this paper, we show the impact of this property on some well known policies. We use a simulation technique that we consider suitable to agree with the real architecture.

**Keywords:** QoS; Heavy-Tailed distribution; Simulation; Task Assignment; Performance; Queuing Theory; Load Sharing.

## 1 Introduction

The explosion of Internet and the World Wide Web has sensibly increased the amount of on-line information and services available for people. The growing information and service demands have placed a dramatic pressure on the Internet infrastructure by the need of advanced Web server systems able to serve a large number of Web requests. Hence, the developers turned to the distributed designs of Web server systems (Web clusters), because of their scalability and cost-effectiveness, instead of one Web server with high performance like mainframe.

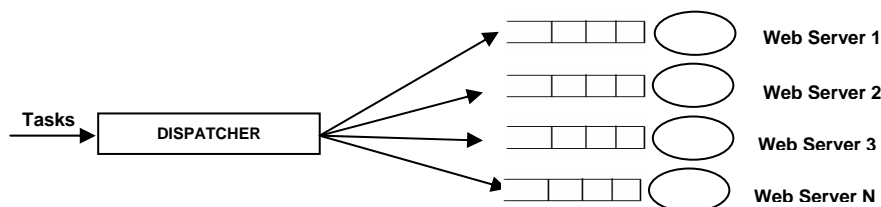


Fig. 2. Distributed Web server model

Fig. 1. Illustrates a typical architecture of a distributed Web server model. The arriving tasks (requests for Web files) to the distributed Web server system must be assigned, by the dispatcher, to one of the Web server to be processed. The Quality of Service (QoS) in such a system focuses on the assignment policy of these tasks in order to minimize their response time and thus, improve the performance of system [1]. The response time of a task consists of the time taken by the approach algorithm to assign a task and the flow time. So, improving the response time tends to minimize the related metrics which are the waiting time, the slowdown and the flow time.

In [2, 3, 4], recent observations show that the task size (service demand) often exhibit heavy-tailed distribution where  $\Pr\{X > x\} \sim x^{-\alpha}$ ,  $0 < \alpha < 2$ , and where, a small fraction (about 3%) of large tasks makes half (50%) of the total load, known also as heavy-tailed property. In this paper, we show the impact of this property on two well known policies for task assignment in distributed Web server systems. We use a statistical simulation technique which agree with a real architecture of such systems.

The remainder of this paper is structured as follows: Section 2 reviews some background information that will be useful to the rest of this paper. A description of the simulation of the two policies for QoS in distributed Web server systems is given in Section 3 and we conclude our work in Section 4.

## 2 Background

In this section, we show the distribution that has been considered to represent traffic experienced by distributed Web server systems and the metrics to evaluate the performances of such systems.

### 2.1 Task sizes distribution (Heavy-tailed distribution)

First, heavy-tailed distribution have been observed in many natural phenomena: like the distribution of people around the world where most places are empty or tiny populated, while there are small numbers of places which are large populated.

The distribution of a random variable  $x$  is heavy-tailed if:

$$\Pr \{X > x\} \sim x^{-\alpha}, \quad 0 < \alpha < 2 \quad (1)$$

The simplest heavy-tailed distribution is the Pareto distribution, with probability density function:

$$f(x) = \alpha k^\alpha x^{-\alpha-1}, \quad \alpha, k > 0 \quad x \geq k \quad (2)$$

and cumulative distribution function

$$F(x) = \Pr \{X > x\} = 1 - (k/x)^\alpha \quad (3)$$

Which  $\alpha$  is the index of variability of the task size and  $k$  the smallest possible task. This means that a random variable that follows a heavy-tailed distribution can be largest with no negligible probability, contrast with the exponential distribution which  $\Pr \{X > x\} \sim 0$ .

Recently this distribution was shown exhibited in the computing systems. Authors in [2, 3, 4] have shown that the task sizes exhibit this distribution where a very small fraction (<1%) of the very largest tasks makes up a large fraction (half) of the total workload. This property is known as the heavy-tailed property. Contrast this with exponential distribution where <1% of the very largest tasks makes up about 5% of the total workload. The heavy-tailed property will be most effective when the variation increases, which means that  $\alpha$  decreases.

### 2.2 Performance metrics

Some metrics were be used to compare the performance of the existing policies for QoS in distributed Web server systems. The waiting time, the slowdown and the flow time are considered.

#### Waiting time

The waiting time is the time from when the task arrives at the system until it begins its service. The waiting time of a task is the total processing time of all tasks that are ahead of the task in the queue. So, the waiting time of a task  $n$ , is:

$$\text{WaitingTime}_n = \sum_{i=1}^{n-1} \text{ProcessingTime}_i \quad (4)$$

Where the *mean processing time* of a Web server  $k$  is:

$$\text{MeanProcessingTime}_k = \frac{\sum_{i=1}^{n_k} \text{TaskSize}_i}{(n_k \times \text{ProcessingCapability}_k)} \quad (5)$$

Where  $\text{ProcessingCapability}_k$  is the processing capacity of the Web server measured as the capacity of CPU, RAM and cache size.

The reducing of the waiting time improves a perceived latency of clients.

### Slowdown

The slowdown of a task is the waiting time divided by its processing time. This metric is so important because minimizing it tends to reduce the variation of small and large tasks. So, small tasks will be never delayed by large ones. The slowdown of a task  $i$  is defined as:

$$\text{SlowDown}_i = \text{WaitingTime}_i / \text{ProcessingTime}_i \quad (6)$$

Hence the *mean slowdown* of  $n$  tasks in the queue is:

$$\text{MeanSlowDown} = \left( \sum_{i=1}^n \text{SlowDown}_i \right) / n \quad (7)$$

### Flow time

The flow time is defined as the sum of the waiting time plus the processing time; known also as the time from task arrival to task departure. The flow time of a task  $i$  is defined as:

$$\text{FlowTime}_i = \text{WaitingTime}_i + \text{ProcessingTime}_i \quad (8)$$

So, the *mean flow time* results as:

$$\text{MeanFlowTime} = \left( \sum_{i=1}^n \text{FlowTime}_i \right) / n \quad (9)$$

## 3 Performance testing

Tasks reaching the system will be assigned, by the dispatcher, to one of the Web servers based on the task assignment approach. At a local Web server, tasks are queued and processed in FIFO order. We have implemented two well known policies for QoS in distributed Web server systems and performed a statistical simulation [5] which generates a synthetic trace file following the Bounded Pareto distribution and calculates the performance metrics. Each set of tasks consists of 5000 tasks keeping the mean service time, at the empirical value of the distribution, equal to 3000 [6, 7, 8]. We summarize these policies as follows:

- **Round-Robin**: tasks are assigned to the participated Web servers in a cyclical fashion. Task  $i$  is assigned to Web server  $i \bmod N$ , where  $N$  is the number of Web servers.
- **SITA-E** [6]: Each Web server has a separate range of task sizes. Tasks are assigned to a Web server according to the task sizes that fall in the Web server's task size range. The approach algorithm aims to equalize the expected load received at each Web server.

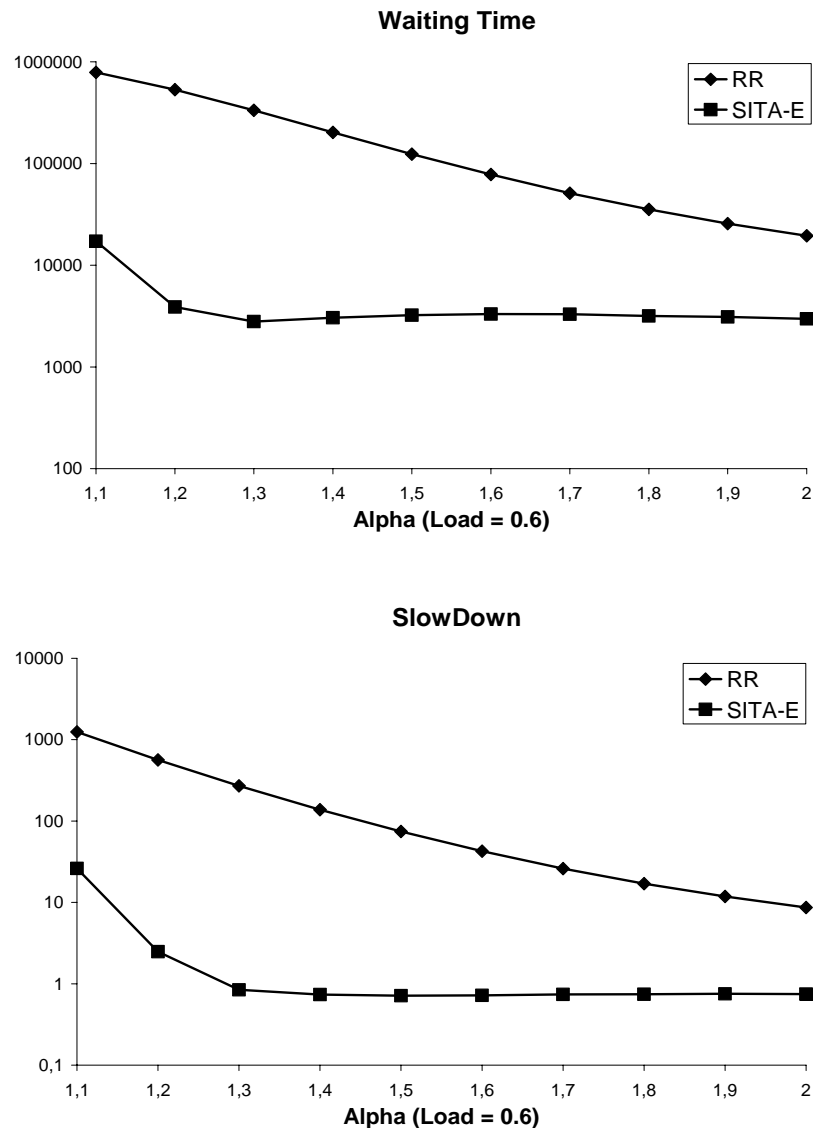
We compare the performances of these task assignment policies using the common metrics. We consider the waiting time, the flow time and the slowdown. The comparisons were performed by varied

the system load  $\rho = \lambda \cdot E(x)/h$  [6], where  $\lambda$  is the arrival rate,  $E(x)$  is the mean service time and  $h$  the number of Web servers.

As shown on figures Fig. 2., Fig. 3. and Fig. 4. the performances of Round Robin are the worst in term of waiting time, slowdown and flow time. This is due to the policy algorithm that tries to equalize the expected number of tasks at each Web server without considering the load of these tasks. So, when the variability increases ( $\alpha$  decreases), Round Robin suffers to balance the load due to the large tasks that will be queued with small ones and thus deteriorate the performance metrics.

On the other side, we show an improvement in performance of SITA-E which, unlike Round Robin, choose a size range for each Web server to equalize its total load.

We show, also, that the performance of SITA-E begins to deteriorate when  $\alpha$  decreases since a great part of tasks are small, so they will be affected to the highest Web servers (due to the size ranges) and, the performances of system will be those of the highest Web servers which are overloaded.



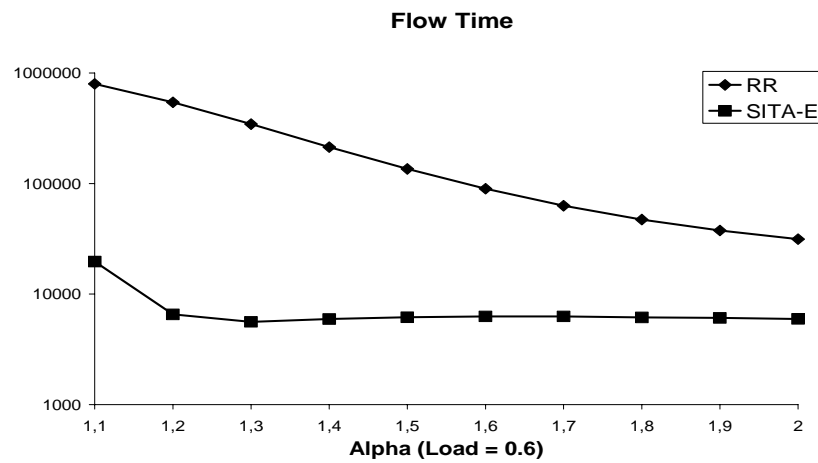
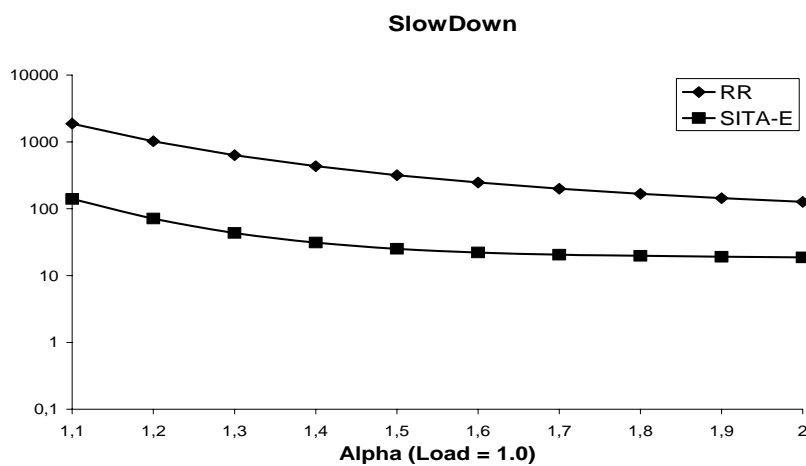
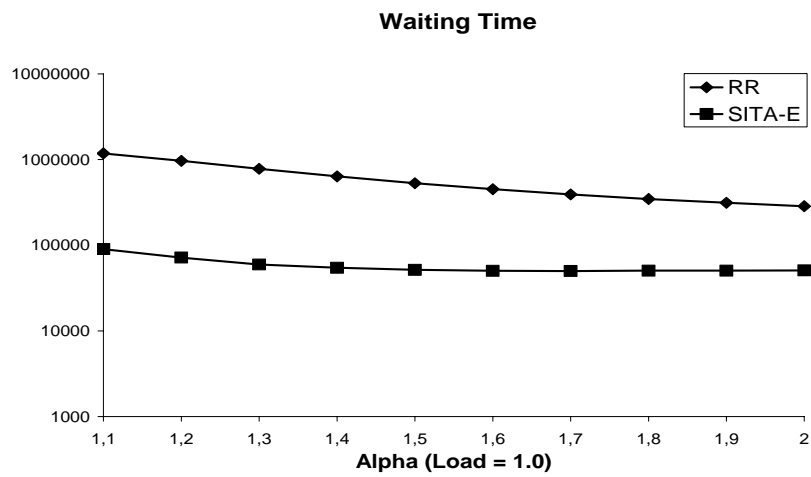


Fig. 3. Performances for under-loaded system



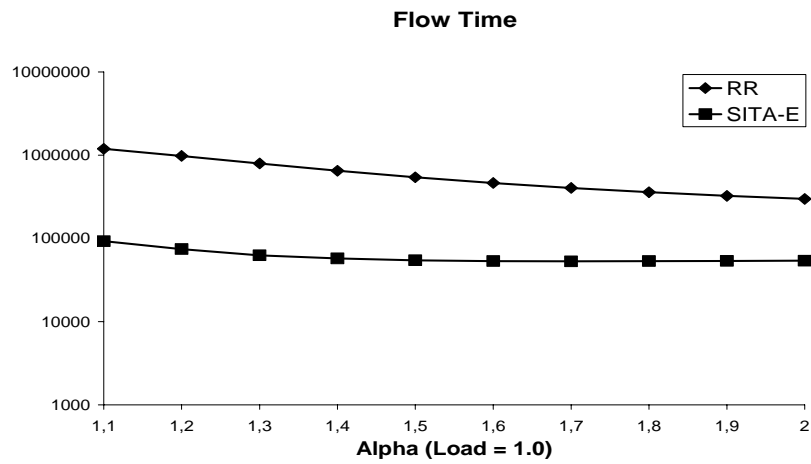
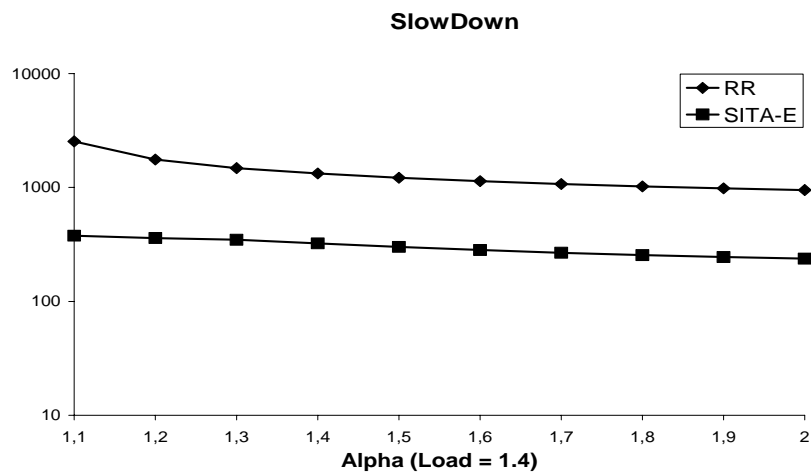
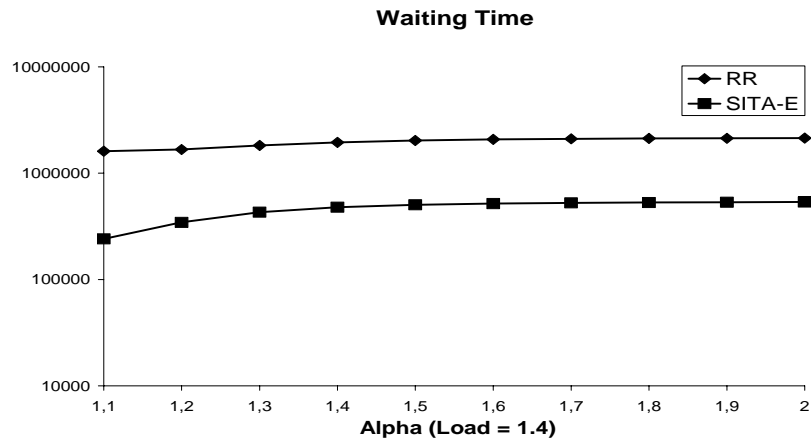


Fig. 3. Performances for loaded system



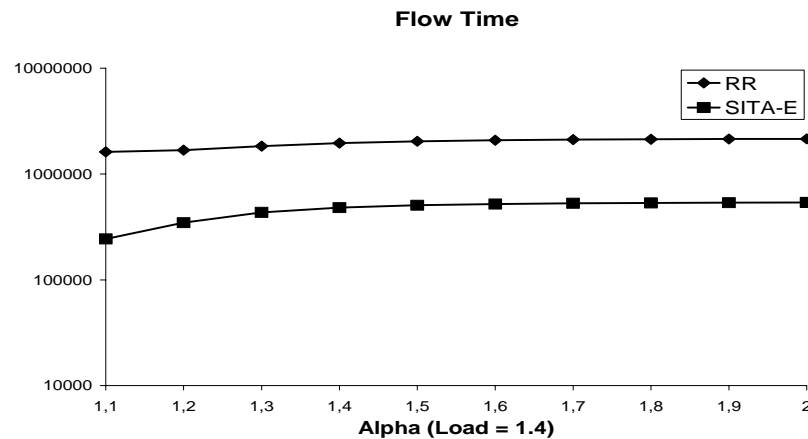


Fig. 4. Performance for over loaded system

#### 4 Conclusion and future works

This paper investigates the task assignment (i.e. QoS) in distributed Web server systems. The task size distribution in such systems follows the heavy-tailed distribution, where a small number of large tasks makes half of the total load.

Performance testing of well known policies for QoS in distributed Web server systems was performed. We have implemented the two policies Round Robin and SIAT-E, and performed a statistical simulation. We showed the impact of the heavy-tailed property on the task assignment policy as the variation of tasks increases.

The problem of QoS in distributed Web server systems have been more studied and, due to the heavy-tailed distribution the proposed solutions remain not suitable to deal with the characteristic of Web traffic and the workload. Our future work will focus on the proposition of new policy that deal with the problem of large tasks, which are difficult to balance among Web servers and delay small tasks when they are in the same queue.

#### References

- Cardellini, V., Casalicchio, E., Colajanni, M. and Tucci, S.: Mechanisms for Quality of Service in Web Clusters. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 37, (2001) 761 - 771.
- Arlitt, M. F. and Williamson, C. L. : Web server workload characterization: The search for invariants. *Proceedings of the 1996 SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, Philadelphia, Pennsylvania, United States, ACM Press, New York, NY, USA (1996) 126-137.
- Crovella, M. E. and Bestavros, A. : Self-similarity in World Wide Web traffic: Evidence and possible causes. *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, Philadelphia, Pennsylvania, United States, ACM Press, New York, NY, USA (1996) 160-169.
- Crovella, M. E., Taqqu, M. S. and Bestavros, A. : Heavy-tailed probability distributions in the world wide web. *A Practical Guide To Heavy Tails*, Chapman & Hall, New York, Birkhauser Boston Inc. Cambridge, MA, USA (1997) 3-25.
- Lizy, J. K., Lieven, E. : *Performance Evaluation and Benchmarking*. CRC Press (2006).
- Harchol-Balter, M., Crovella, M. and Murta, C. : On Choosing a Task Assignment Policy for a Distributed Server System. *Journal of Parallel and Distributed Computing*, 59, (1999) 204-228.
- Fu, B., Broberg, J. and Tari, Z. : Task Assignment Strategy for Overloaded Systems. *Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)*, IEEE Computer Society, Washington, DC, USA (2003) 1530-1546.
- Fu, B. and Tari, Z. : A Dynamic Load Distribution Strategy for Systems under High Task Variation and Heavy Traffic. *Proceedings of ACM Symposium on Applied Computing*, Melbourne, Florida, ACM Press, New York, NY, USA (2003) 1031-1037.

## Application des Métriques de Similarité Sémantique pour la Composition des Services Web

Hadjila Fethallah<sup>1</sup>, Chikh Mohammed Amine<sup>1</sup>, Belabed Amine<sup>1</sup>,  
Djabari Belkhir<sup>1</sup> et Khater Maamar<sup>2</sup>

<sup>1</sup> Université de Abou Bekr Belkaid Tlemcen  
{f\_hadjila, mea\_chikh}@mail.univ-tlemcen.dz  
{amine\_inf, belkheir\_d}@yahoo.fr

<sup>2</sup> Université Djillali Liabes SBA  
khater.maamar@hotmail.fr

**Résumé.** La composition sémantique des services web joue un rôle très important dans l'échange B2B et B2C. En effet la composition permet l'épargne des efforts et des coûts destinés à la construction des nouvelles applications, elle permet aussi la satisfaction des besoins des utilisateurs en cas d'absence d'un résultat direct (sans composition). Dans cet article nous proposons une approche de composition qui se base sur les métriques (distances) sémantiques de l'ontologie wordnet à fin de calculer la similarité entre deux services à composer. Nous utilisons aussi d'autres critères syntaxiques pour améliorer la qualité globale des résultats.

**Mots clés:** Composition de services, Services web, Mesures de similarité, Wordnet, Ontologies, Systèmes multiagents, AI Planning.

### 1 Introduction

Les services web peuvent être définis comme des applications modulaires, auto descriptibles, auto contenues, et accessibles à travers l'internet [7]. Plusieurs standards ont été définis tels que WSDL, UDDI, SOAP [6] afin de permettre la description, la découverte, et l'invocation des services (respectivement).

La prolifération du nombre des services web, entraîne une complexité dans leur recherche et leur composition. Dans un environnement ouvert et dynamique, tel que Internet, on assiste de plus en plus à un besoin de composition sémantique de services, en effet, la prise en charge la sémantique durant le matching des services, va minimiser les fausses réponses, et améliorer nettement la qualité globale des résultats.

Pour résoudre ce problème, différents efforts ont été dépensés dans la littérature [8], [12]. Ces approches diffèrent par leur degré d'automatisation (manuelles, interactives, automatiques), par leur prise en charge de la sémantique (basées sur les ontologies ou non), et par leur caractère dynamique vs statique. Pour gérer toutes ces difficultés, notre contribution est articulée sur :

- les systèmes multiagents (SMA), qui reposent sur les concepts de l'I.A.D [4].
- la similarité syntaxico-sémantique de la base de données lexicale wordnet. En effet, deux services s1 et s2, dont l'entrée du premier est compatible avec la sortie du deuxième (au sens de la mesure de similarité) peuvent être composés afin de créer un service à valeur ajoutée. Le SMA développé est constitué :
  - d'un agent client qui initie le processus de composition, en émettant une requête contenant les i/o du service (inputs, outputs)
  - d'un agent médiateur faisant le lien entre les agents fournisseurs et le client,
  - d'un ou plusieurs agents fournisseurs de services qui réalisent l'inférence sémantique des plans, et les retransmettent au médiateur.

Le reste de l'article est structuré comme suit : La deuxième section introduit la notion d'agents et de SMA, la troisième section présente le thésaurus wordnet, la quatrième section présente un ensemble de mesures de similarité. La cinquième section présente l'approche développée, la sixième section propose un survol sur les travaux effectués dans la composition de services. Et enfin la septième section clôture le travail par une conclusion et présente les perspectives.

### 2 Les Agents Intelligents:

L'intelligence artificielle distribuée (IAD) s'intéresse aux actes intelligents qui résultent de l'activité coopérative de plusieurs agents. Pour résoudre un problème complexe, il est nécessaire de le diviser en



sous problèmes et de les affecter aux composants (agents) du système. Plusieurs définitions ont été proposées pour éclaircir la notion d'agent, on présente par la suite la plus importante.

Un agent est une entité autonome, réelle ou abstraite, qui est capable d'agir sur elle-même et sur son environnement, qui, dans un univers multi-agents, peut communiquer avec d'autres agents, et dont le comportement est la conséquence de ses observations, de ses connaissances et des interactions avec les autres agents [9].

### 3 Wordnet

WordNet [10], est une ontologie lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. Son but est de recenser, classer et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Elle peut être considérée comme un réseau sémantique où chaque nœud représente un concept du monde réel. Chaque nœud appelé synset, est composé d'un ensemble de synonymes qui représentent le même concept. Les synsets sont reliés par des arcs qui décrivent les relations linguistiques entre eux, elles peuvent indiquer soit l'hypéronymie, l'hyponymie, la méronymie, l'antonymie...

### 4 Les Mesures de Similarités

Le problème de similarité a été largement étudié dans les domaines de l'intelligence artificielle, la psychologie, la linguistique...etc. pour notre cas, la similarité entre les concepts d'ontologie, peut être calculée en se basant sur le nombre d'arcs intermédiaires séparant les deux nœuds, plus le nombre d'arcs intermédiaires est petit plus la similarité est grande. Les autres méthodes (les approches basées sur les nœuds), utilisent les mesures de contenu d'informations pour déterminer la similarité conceptuelle. Les sections 4.1, 4.2 présentent deux exemples appartenant à la première classe d'approches, et la section 4.3 présente une mesure d'ordre syntaxique.

#### 4.1 Mesure de Wu-Palmer

Dans un thesaurus de concepts tel que wordnet, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine [19]. La similarité entre C1 et C2 est :

$$\text{Sim1}(C1, C2) = (2 * N) / (N1 + N2 + 2 * N)$$

Où N est la profondeur du subsumant le plus spécifique de C1 et C2 (noté C) par rapport à la racine.

Ni est la profondeur de Ci par rapport à C (avec i appartenant à {1,2}). Cette mesure a l'avantage d'être simple à implémenter.

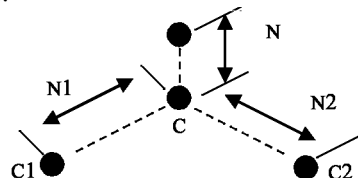


Fig. 1. La distance de (Wu & Palmer).

#### 4.2 Mesure à Base de Milestone:

Cette mesure est un élément clé de notre approche de composition, elle s'inspire des travaux cités dans [20]. La similarité entre deux concepts dépend de la distance  $d(c1, c2)$  qui sépare leur positions dans l'hérarchie, formellement :

$$\text{Sim2}(c1, c2) = 1 - d(c1, c2)$$

Chaque noeud possède une valeur appelée milestone [20], évaluée comme suit:  $\text{milestone}(n) = (1/2)^{kl(n)}$ , avec:

$k$  : facteur supérieur à 1. Il sert à diminuer la valeur en fonction de la profondeur du noeud dans l'hierarchie.

$l(n)$  : représente la profondeur du noeud  $n$  dans l'hierarchie.

La distance est calculée comme suit:

$d(c1, c2) = d(c1, c) + d(c2, c)$  et  $d(c, c) = \text{milestone}(c) - \text{milestone}(c)$  Intuitivement la distance compte le nombre d'arcs intermédiaires en passant par leur plus petit subsumant noté  $c$ .

Ce modèle implique deux suppositions: La différence sémantique entre les concepts du niveau supérieur est plus grande que ceux du niveau inférieur. En d'autres mots:

02 concepts généraux sont moins similaires que deux concepts spécialisés en plus la distance entre frères est plus grande de la distance entre un père et son fils.

#### 4.3 Mesures Syntaxiques Basées sur les Ngrams :

Un  $n$ -gram [1], est une séquence de  $n$  caractères consécutifs. L'ensemble des  $n$ -grams (en général  $n$  prend les valeurs 2 ou 3) qu'on peut générer est le résultat qu'on obtient en déplaçant une fenêtre de  $n$  cases sur le corps de texte.

Pour appliquer les ngrams sur wordnet on calcule tous les  $n$ -grams des deux termes à comparer, plus particulièrement, on va prendre  $n=3$  et on compte les occurrences de chaque 3grams pour chaque terme. Et ensuite on applique la distance 'cosine' que l'on a adapté à partir du domaine de recherche d'informations [2]:

$$\text{Cosine}(a1, b1) = \frac{(\sum_{w \in a1 \cap b1} (\text{nb\_oc}(w, a1) \times \text{nb\_oc}(w, b1)))}{\sqrt{(\sum_{w \in a1} \text{nb\_oc}^2(w, a1)) \times (\sum_{w \in b1} \text{nb\_oc}^2(w, b1))}}$$

Fig. 2. La mesure cosine

Avec  $a1, b1$  : les deux termes à comparer,  $w$  est un 3gram contenu dans un terme ( $a1$  ou  $b1$ ).

$\text{Nb\_oc}(w, a1)$  : le nombre d'occurrence du 3gram  $w$  dans le terme  $a1$ . Même chose pour

$\text{Nb\_oc}(w, b1)$ .

Si les synsets de  $a1$  et de  $b1$  contiennent plusieurs termes alors on va prendre la mesure (cosine) la plus grande entre les deux ensembles de synsets, ie :

Si  $\text{synset\_a1} = \{a1, a2, \dots, an\}$  et  $\text{synset\_b1} = \{b1, b2, \dots, bm\}$  alors :

$$\text{Cosine}(a1, b1) = \max\{\text{cosine}(ai, bj)\}$$

Avec  $ai \in \text{synset\_a1}$  et  $bj \in \text{synset\_b1}$ . Plus le nombre des 3grams communs entre les deux termes à comparer est grand plus la mesure de similarité est élevée.

#### 5 Approche Proposée

Pour faciliter la compréhension de l'approche on considère l'exemple suivant qui concerne la vente en ligne des ouvrages: nous avons par exemple trois services  $s1, s2, s3$ .  $s1$  fournit le code ISBN d'un ouvrage à partir du titre,  $s2$  donne le prix en euro à partir du code ISBN,  $s3$  convertit le prix mesuré en euro en dinar algérien.

En supposant qu'un utilisateur cherche un service donnant le prix en dinar algérien d'un ouvrage à partir du titre, il est évident que cette requête n'est pas disponible chez les fournisseurs, mais elle peut être composée, dynamiquement par un ou plusieurs agents logiciels. Le résultat d'une telle composition satisfera complètement les besoins de l'utilisateur.

### 5.1 Mesure de Similarité Proposée

Le processus de matching de la requête avec les services existants utilise la mesure hybride suivante:

$$\text{Sim\_hybride}(c1,c2)=w*\text{sim2}(c1,c2)+(1-w)*\text{cosine}(c1,c2).$$

**Fig . 3.** La mesure utilisée

Avec : c1, c2 des concepts de l'ontologie wordnet.

sim2 : la mesure présentée dans section 4.2.

Et cosine est une mesure syntaxico-sémantique qui utilise les 3grams et le contenu des synsets (les concepts de l'ontologie wordnet), elle est présentée dans la figure 2.

Le paramètre w est mis à 0.8 (valeur par défaut), mais il peut être déterminé expérimentalement (selon les critères de rappel et de précision). L'expérimentation fera l'objet d'un travail futur.

Pour composer deux services s1 et s2, qui possèdent des entrées et des sorties complexes, (ie des structures de données composées de plusieurs champs) :

- on crée une matrice de correspondance, qui contient les Ii dans ses lignes, et les Oj dans ses colonnes. Avec :

Ii : une entrée (input) de s1

Oj : une sortie de s2.

- Ensuite on calcule le score de similarité (matching), pour toutes les combinaisons (Ii,Oj) à travers la formule de la figure 3. La table 1 montre un exemple. - On classe les scores par ordre décroissant.

- On retient les couples ayant les scores les plus élevés, tout en respectant les contraintes suivantes:

Une entrée Ii n'a qu'un seul correspondant dans la liste des Oj

De même une sortie Oj n'a qu'un seul correspondant dans la liste des Ii

**Table 1.** Exemple d'un matching de structures de données.

S2 ouput s	O1	O2	O3
I1	0.75	0.5	0.6
I2	0.9	0.2	0.3

Pour l'exemple de la table1, on retient les éléments (I2, O1, 0.9) et (I1, O3, 0.6). L'élément (I1, O1, 0.75) ne peut être retenu, parce que O1 est déjà pris.

- dans la dernière étape on calcule le score global comme suit :

Score-glob=(sous scores)/nombre de Ii.

Pour l'exemple précédent le score global sera : Score-glob =(0.9+0.6)/2=0.75. L'idée de base de l'approche proposée est d'utiliser la coopération multiagent et la mesure hybride (figure 3), pour former des plans dynamiques de services. Plus particulièrement on a adopté les constituants suivants :

L'agent client (a.c): c'est l'agent qui va initier l'interrogation (recherche) sémantique, il émet au médiateur, une requête constituée d'entrées, de sorties, et une référence sur le domaine de services dont on a besoin (par exemple la vente en ligne des ouvrages), et d'un taux de matching expliqué ultérieurement.

L'agent médiateur (a.m): il enregistre les agents fournisseurs entrant au système en notant les informations suivantes : nom agent, localisation (adresse), domaine de son travail par exemple tourisme, vente d'ouvrages...

Son rôle est de :

- lancer périodiquement un processus d'apprentissage distribué sur les agents fournisseurs, afin de composer sémantiquement des éventuels plans de services.

- router les requêtes des clients aux agents fournisseurs adéquats (qui ont le même domaine), et de récupérer les réponses (services offerts) avant de les envoyer au client.

Les agents fournisseurs (a.f) : ces agents possèdent des descriptions sémantiques sur les services offerts. Ils sont responsables du matching et de la composition à base de similarité sémantique (voir les sections 5.2 et 5.3).

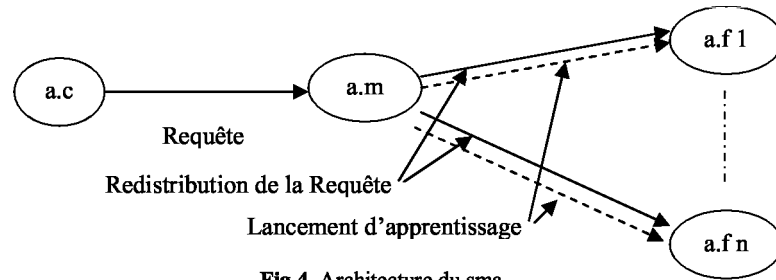


Fig.4. Architecture du sma

## 5.2 Composition Sémantique des Services

La composition se déroule selon le scénario suivant :

1- le médiateur diffuse périodiquement un message contenant la liste des agents fournisseurs (a.f) qui doivent faire l'apprentissage des services composites (ces a.f appartiennent au même domaine).

2- Chaque a.f détient une table qui décrit la sémantique des services, elle est constituée des champs suivants : input, output, similarité intra-service, nombre de sous services (les composants), noms des sous services, liste des clés-uddi associées aux sous services.

La similarité intra-service notée Sim est calculée comme suit :

- ✓ Elle vaut 1 pour un service atomique.
- ✓ Pour un service composé de 02 sous services s1, s2 elle est calculé à l'aide de la mesure présentée par la figure 3, elle utilise l'entrée de s1 et la sortie de s2.
- ✓ Pour un service composé de plusieurs sous services, elle vaut la moyenne des similarités de chaque paire de sous services.

3- Les agents de la même liste échangent leurs tables de description de services notées t.d.s (de la même manière que le protocole RIP fonctionne)

4- Chaque agent k compare à l'aide de la mesure de la figure 3, l'entrée de chaque Slj (service numéro l appartenant à l'agent j), avec la sortie de Ski (service numéro i appartenant à l'agent k). le Slj est récupéré des t.d.s reçus.

5- Si la similarité  $\text{sim}(\text{Skj}, \text{Sli})$  seuil fixé au départ du système (par exemple 80%) alors ajouter le service composite modélisé par le 6-uplet suivant (Input Ski, output Slj,  $\text{sim}(\text{Skj}, \text{Sli})$ , 2, [Ski, Slj], [clé Ski, clé Slj]), à la table t.d.s de l'agent k.

Il est à noter que le seuil peut être changé par la suite.

6- Chaque agent k essaye de comparer uniquement les services de sa t.d.s en regardant s'il y a une similarité sémantique acceptable (selon le seuil), entre chaque couple de services (Skj, Ski). La nouvelle similarité du service composite est la moyenne du triplet :  $\text{sim}(\text{Skj}, \text{Ski})$ , Sim de Ski, et Sim de Skj.

7- Ce processus est répété tant qu'il y a des services composables.

## 5.3 Matching Sémantique des Services

La requête d'un agent client est composée d'un input, d'un output, et d'un taux de matching. Elle parvient aux fournisseurs a.f à travers l'agent médiateur. Le taux de matching mesure le degré de similarité entre la requête et le service offert.

Le matching s'obtient à travers les similarités suivantes (au sens de la figure 3) :

- ✓ Une similarité entre Input\_requête et Input\_service notée  $\text{sim\_input}$ .
- ✓ Une similarité entre output\_requête et output\_service notée  $\text{sim\_output}$ . On prend la moyenne de  $\text{sim\_input}$ ,  $\text{sim\_output}$  et de la Sim du service lui-même (récupérée de la t.d.s).
- ✓  $\text{similarité\_matching} = (\text{sim\_input} + \text{sim\_output} + \text{Sim du service offert})/3$ .

On retient uniquement les services présentant une similarité de matching compatible au taux de matching de l'utilisateur ie  $\text{similarité\_matching} \geq \text{taux de matching}$ . Après, chaque agent k envoie indirectement ses résultats au client.

## 6 Etat de L'art

Les approches de composition de services web peuvent être regroupées en deux catégories. La première se base sur les workflows, et permet d'adapter les méthodes d'orchestration et de

chorégraphie afin de créer des processus métiers. On peut citer par exemples les références [3], [13] pour ce type d'approches. Dans la deuxième catégorie, la composition est vue comme la génération automatique d'un plan d'exécution des services Web (AI Planning). Les références [14],[15],[16],[18] montrent différents langages faisant partie de ce paradigme, les références [5], [11], [17] présentent une version distribuée du AI planning en utilisant les SMA.

## 7 Conclusion

Dans ce papier nous avons présenté un système multiagent pour la composition dynamique des services web sémantiques. Chaque service possède une description sémantique contenue dans une table appelée t.d.s. Le système proposé utilise la mesure de similarité hybride, qui combine les aspects sémantiques (la liste des synonymes), et les aspects syntaxiques (les mesures basées sur les ngrams). Nous avons montré aussi un protocole d'interaction qui regroupe trois types d'agents : l'agent client, l'agent médiateur, et les agents fournisseurs. Ces derniers ont un rôle pivot dans le processus d'apprentissage des services composites. Comme perspectives à ce travail nous suggérons d'autres critères de composition (par exemple la qualité de service Qos).

## Références

1. R. C. Angell, G. E. Freund, "Automatic spelling correction using a trigram similarity measure", *Information Processing and Management* 19(4), (1983). 255-161.
2. Baeza-Yates, R., and Ribeiro-Neto, B. d. A. *Modern Information Retrieval*, Addison-Wesley, Reading, MA, (1999).
3. BEA, IBM, Microsoft. *Business Process Execution Language for Web Services (BPEL4WS)*. <http://xml.coverpages.org/bpel4ws.html> (2003).
4. Bond, A and Gasser, L. *Readings in Distributed Artificial Intelligence*. San Mateo, California, Morgan Kaufman Publishers. (1988).
5. Bourdon, J. *Multi-agent systems for the automatic composition of semantic web services in dynamic environments*, Master's thesis, ENSM Saint- Etienne (2007).
6. Curbera, F., Duftler, M. Khalaf, R., Nagy, W., Mukhi, N. and Weerawarana, S. "Unraveling the Web Services Web: An Introduction to SOAP, WSDL, and UDDI." *IEEE Internet Computing*, 6(2). (2002).
7. Curbera, F., Nagy, W., and Weerawarana, S. "Web Services: Why and How." *Workshop on Object-Oriented Web Services – OOPSLA*, Tampa, Florida, USA. (2001).
8. Dustdar, S. and Schreiner, W. "A survey on web services composition", *Int. J. Web and Grid Services*, Vol. 1, No. 1, pp.1–30. (2005).
9. Ferber, J. *Les systèmes multi-agents: Vers une intelligence collective*. InterEditions. (1995).
10. Fellbaum, C. *WORDNET. An Electronic Lexical Database*. In *The MIT Press* (1998).
11. Guitton, J., Humbert F. *Planification multi-agent pour la composition dynamique de services Web* thèse de Master, Université Joseph Fourier – Grenoble I. (2006).
12. Hadjila, F., M. Chikh. *La composition des services web sémantique : une approche à base de distance sémantique*. Siie'2008 hammamet, Tunis. (2008).
13. Laukkanen, M., and Helin, H. *Composing workflows of semantic web services*. In *Proceedings of the Workshop on Web-Services and Agent-based Engineering*. (2003).
14. McDermott, D. *Estimated-regression planning for interactions with Web services*. In *Proceedings of the 6th International Conference on AI Planning and Scheduling*, Toulouse, France. (2002).
15. McIlraith, S., and Son, T. *Adapting Golog for composition of semantic web services*. In *Proceedings of the 8th International Conference on Knowledge Representation and Reasoning (KR '02)* Toulouse, France, Morgan Kaufmann Publishers, pp. 482–493. (2002).
16. Medjahed, B., Bouguettaya, A., and Elmagarmid, A.K. *Composing Web Services on the Semantic Web*. *The VLDB Journal — The International Journal on Very Large Data Bases*, Volume 12 Issue 4. (2003).
17. Pellier, D. *Modele dialectique pour la synthese des plans* Phd thesis UJF- grenoble. (2005).
18. Rao, J., Kungas, P., and Matskin, M. *Application of linear logic to web service composition*. In *Proceeding of the 1st International Conference on Web Services (Las Vegas, USA)*. (2003).
19. Wu, Z., and Palmer, M. "verb semantics and lexical selection" In *Proceedings of the 32nd annual meeting of the association for computational linguistics*, New Mexico. (1994).
20. Zhong, J., Zhu, H., Li, J. Yu, Y.: *Conceptual Graph Matching for Semantic Search*. In: *Proceedings of the 2002 International Conference on Computational Science*, (2002).

## Editeur SCORM en web services conforme à la norme IMS GWS « Instructional Management Systems General Web Services »

Faiza Medjek <sup>1</sup>, Aouaouche El-Maouhab <sup>1</sup>, Radia Adjrad <sup>1</sup> et Nahla Ghehria <sup>1</sup>,

<sup>1</sup> Centre de Recherche sur l'Information Scientifique et Technique (CERIST), Division Technologies des Réseaux d'Information (DTRI)

03 rue frères Aissiou, Benaknoun, Alger

Medjek-f@dtri.cerist.dz, elmaouhab@wissal.dz, adjradradia@hotmail.com, mmorpg715@hotmail.com

**Résumé.** Un éditeur de cours conforme SCORM (Sharable Content Object Reference Model ou Model de Référence d'Objet de Contenu Partageable) permet l'ajout, la modification, la suppression et le regroupement des objets pédagogiques en contenu pédagogique (cours, chapitre,...) partageable, réutilisable et qui permet l'interopérabilité entre différents LMS (Learning Management System ou Système de Gestion de l'Apprentissage).

La norme IMS General Web Services permet l'interopérabilité dans la couche application, en particulier, la description des comportements exposés par l'intermédiaire des services web.

Dans le présent travail, nous proposons de décrire comment découper les fonctionnalités d'un éditeur SCORM déjà réalisé en un certain nombre de fonctions, qui peuvent être implémentées séparément sous la forme d'application autonomes (stand-alone) ou sous la forme de e-services. Pour cela nous suivons les directives de la spécification IMS GWS. La mise à disposition de ces services permet la réutilisation des fonctionnalités dans une plate-forme *e-Learning*.

**Mots-Clés:** Web services, e-learning, LMS, IMS GWS, SCORM

## 1 Introduction

Ces dernières années ont vu Internet prendre de plus en plus d'importance et finalement devenir la plate-forme universelle de diffusion d'informations, avec un essor de la communication, de la collaboration et de l'enseignement à distance. La vue sur l'enseignement prend un nouveau visage, nous assistons à un nouveau besoin, de l'enseignement traditionnel dispensé dans des salles de cours, nous migrons vers une demande en temps réel de l'information.

Parallèlement à cette évolution sont apparus les LMS (Learning Management Systems) qui avaient pour objectif d'aider les entreprises à gérer leur formation. Mais l'application des normes dans un contexte éducatif est parmi les enjeux majeurs de cette nouvelle évolution. Les normes dans le domaine du e-learning, parmi lesquelles on cite la norme SCORM, se focalisent sur la réutilisation des objets pédagogiques mais ne concernent pas la réutilisation des fonctionnalités des éditeurs.

Dans le présent article nous proposons de décrire la mise en place d'une collection de services Web, qui sont le résultat d'un découpage des fonctionnalités de base d'un éditeur de cours conforme à la norme SCORM. La mise à disposition de ces services permet la réutilisation des fonctionnalités d'un éditeur dans une plate-forme *e-Learning*. Pour pouvoir réaliser cet ensemble normalisé de services web, nous avons opté pour l'implémentation de la spécification IMS General Web Services.

## 2 Les LMS et la normalisation

L'application des Technologies de l'Information et de la Communication (TIC) au domaine de la formation a conduit à la création de cette nouvelle réalité appelée *e-learning*. Décrit comme le mariage du multimédia (son, image, texte) et de l'Internet (diffusion on line, interactivité), il permet de cumuler les avantages de la personnalisation et de la démultiplication et ne présente que peu d'inconvénients. Il représente, cependant, un changement fort des habitudes de travail des utilisateurs et des formateurs.

Les plates formes d'enseignement à distance sont très nombreuses et offrent un choix très varié de services pour les usagers. Elles sont aujourd'hui communément appelés « Systèmes de Gestion de l'Apprentissage » (ou Learning Management Systems).

En effet, le LMS est une solution axée sur l'intégration, la diffusion et la gestion de contenu pour la formation à distance. Il désigne une suite de fonctionnalités conçues pour assurer la présentation, le suivi, la production de rapports et la gestion d'un contenu d'apprentissage, des progrès des élèves et de leurs interactions. Cette expression peut aussi bien désigner des systèmes de gestion de cours très simples que des environnements distribués très complexes. [1]

Dans le contexte de SCORM, on prévoit que les réalisations varieront considérablement. Le modèle de référence SCORM est axé sur les principaux points interfaces entre le contenu et l'environnement du LMS. Il ne spécifie pas les fonctionnalités et les capacités particulières offertes par ce dernier. SCORM partage le contenu pédagogique en asset (actif) : objet de contenu le plus élémentaire (media, document, image, page web, ...), SCO : objet de contenu partageable qui est constitué d'un ou plusieurs assets et CP : agrégation de contenu qui représente le package qui contient l'ensemble des SCO, leur métadonnées et le fichier «imsmanifest» pour la description de l'ensemble [9].

Dans le contexte de SCORM, le sigle LMS renvoie à un environnement à base de serveur qui héberge la logique de gestion de la présentation du contenu d'apprentissage aux élèves. En d'autres termes, dans ce modèle, le LMS est en mesure de déterminer ce qu'il doit présenter et à quel moment et de suivre les progrès des élèves à travers le contenu d'apprentissage. [1]

Toutefois, SCORM ne résout pas tous les problèmes techniques qui doivent être surmontés afin de créer une économie d'objets d'enseignement robuste [3] [7] [8]. Les efforts de standardisation dans ce domaine se focalisent sur la réutilisation des documents pédagogiques [3], mais ne concernent pas la réutilisation des fonctionnalités des LMSs. Ce qui nous a menés à l'intégration de la norme **IMS General Web Services**.

### 3 Les Web Services

Un service web est un composant logiciel autonome possédant un URI (Uniform Resource Identifier) unique, et fonctionnant sur Internet. C'est une application qui se sert des protocoles de communication tel que http pour faire transiter des données dans un format XML (Extensible Markup Language). En clair, c'est un processus permettant à des applications de communiquer entre elles [2] [4].

Les avantages des architectures services Web ont été bien utilisés dans le domaine des applications de type Business-To-Business (B2B) pour l'intégration d'applications d'entreprise et même dans des scénarios de type Business-To-Customer (B2C).

Les services web sont des processus métiers ou des données accessibles via Internet par n'importe quel client. Ils permettent aux applications d'interagir entre elles via le Web. On peut envisager grâce aux services web de segmenter les applications en plusieurs composants ou services partagés, qui peuvent résider sur des machines différentes et de nature complètement hétérogènes. La communication entre les différents acteurs se fait via les langages XML et le protocole http [2] [4].

L'architecture des services web s'est imposée grâce à sa simplicité, à sa lisibilité et à ses fondements normalisés. Le concept des services web s'appuie sur trois éléments essentiels : [2] [4]

- SOAP (Simple Object Access Protocol): un protocole léger fondé sur XML est utilisé pour échanger des informations.

- WSDL (Web Services Description Language): un langage qui décrit les paramètres du service Web, toujours basé sur XML.

- UDDI (Universal Description Discovery & Integration): une architecture répartie détient la description des services fournis.

### 4 IMS

IMS Global Learning Consortium Inc est un consortium regroupant des acteurs du monde éducatif, du privé et des organisations gouvernementales. L'appellation "IMS" vient du nom du projet d'origine "Instructional Management Systems". En 1997, l'IMS a commencé comme étant un projet dans l'initiative d'étude nationale d'infrastructure d'EDUCAUSE aux Etats-Unis. Cette initiative cherche à transformer l'éducation par l'utilisation des nouvelles technologies et son objectif est d'obtenir une large adhésion à des spécifications techniques relatives à la gestion d'outils et de contenus pédagogiques sur Internet. IMS a agi comme un catalyseur pour le développement d'un corpus de logiciels éducatifs, la création d'une

infrastructure en ligne pour gérer l'accès aux matériaux et aux environnements éducatifs, la simplification des travaux éducatifs en groupe et l'évaluation des connaissances et des savoir-faire acquis [4].

IMS a deux objectifs principaux:

- Définir des spécifications techniques pour l'interopérabilité d'application et de service de e-learning;
- Promouvoir l'intégration de ces spécifications dans des produits et services à travers le monde.

L'IMS a développé et a libéré beaucoup de spécifications par lesquelles nous citons:

Tuteur

IMS Access For All Meta-Data;  
IMS Content Packaging;  
IMS Enterprise Services;  
IMS Learning Design;  
IMS Meta-Data;  
IMS Question and Test Interoperability.  
*IMS General Web Services.*

## 5 IMS General Web Services

La plupart des solutions existantes, qui peuvent gérer d'une manière dynamique les contenus pédagogiques, se situent toujours dans le cadre des systèmes centralisés et emploient des formats propriétaires. Cela pose d'autres problèmes, dont celui de l'interopérabilité des systèmes e-Learning. La spécification d'IMS GWS favorise l'interopérabilité à travers des spécifications basées sur des services Web sur différentes plateformes [4].

L'IMS General Web Services est un ensemble de spécifications basé sur les services web. Son objectif est de fournir un cadre pour les équipes de projet de guidage et de permettre l'utilisation des services Web en tant qu'élément de développement de l'étude globale des spécifications du consortium d'IMS (IMS/GLC). L'utilisation des GWS fournit une méthodologie et un profil d'applications qui ont les caractéristiques suivantes: [4]

- **Interopérabilité** : Permettre d'identifier les mécanismes et les normes qui favorisent l'interopérabilité entre les implémentations (réalisations) des services Web à travers différents logiciels et systèmes d'exploitation.
- **Efficacité** : Permettre d'aider d'autres spécifications d'IMS et d'évaluer des protocoles des services Web.
- **Uniformité**: Les spécifications d'IMS GWS permettent de faciliter l'implémentation d'approches uniformes (cohérentes) à l'implémentation des protocoles des services web à travers des projets et des spécifications d'IMS.
- **Flexibilité** : Les spécifications d'IMS GWS sont assez flexibles pour s'adapter aux évolutions des protocoles de service Web tels que « SOAP » et travailler avec une variété de méthodes obligatoires pour des services Web tels que « WSDL ».
- **Aspect pratique** : Les spécifications d'IMS GWS cherchent à faciliter au fournisseur d'implémenter les solutions d'IMS/GLC basées sur les services Web et l'interopérabilité à travers des plateformes et des protocoles de service Web.

D'une perspective technique, les spécifications IMS GWS sont produites pour s'assurer que tous les services définis par IMS/GLC emploient une infrastructure commune et ainsi compatible, d'échange de messages. La figure 1 est une représentation schématique du rapport entre des spécifications d'IMS/GLC et l'IMS GWS. [4]

Sur la figure 1, ci-dessous, les parties principales de cette infrastructure sont : [4]



- **IMS XML Components:** Les composants de l'application et de services communs qui sont combinés pour créer le système e-Learning exigé. On suppose que ces composants échangent l'information sous forme de documents XML ;
- **XML-based Context:** Les documents XML sont transformés en messages XML qui sont alors tracés sur l'infrastructure commune de transmission de messages XML qui est conçue pour soutenir le transfert de données fiable bout à bout exigé de services par exemple, datagramme, pour éditer et souscrire,...etc. L'IMS GWS définit comment le service-description est transformée en équivalent de WSDL ;
- **XML-based Envelope:** Le système commun de transmission de messages XML peut être soutenu en utilisant plusieurs types d'encapsulation d'enveloppe XML c'est à dire SOAP/SOAP avec des attachements de message,...etc. L'IMS GWS définit la structure de message, la chorégraphie et d'autres considérations de transmission de messages ;
- **Generic transport:** L'enveloppe est alors transportée à travers le réseau en utilisant un protocole de transfert de fichier (FTP) approprié ou en utilisant le protocole de transfert hypertexte (HTTP) ;
- **Communications network:** C'est le réseau informatique réel qui est employé pour transporter physiquement les données d'un système à un autre. Ceci sera basé sur la combinaison omniprésente du Protocole TCP/IP fournissant une connexion entre le câble et les réseaux sans fil.

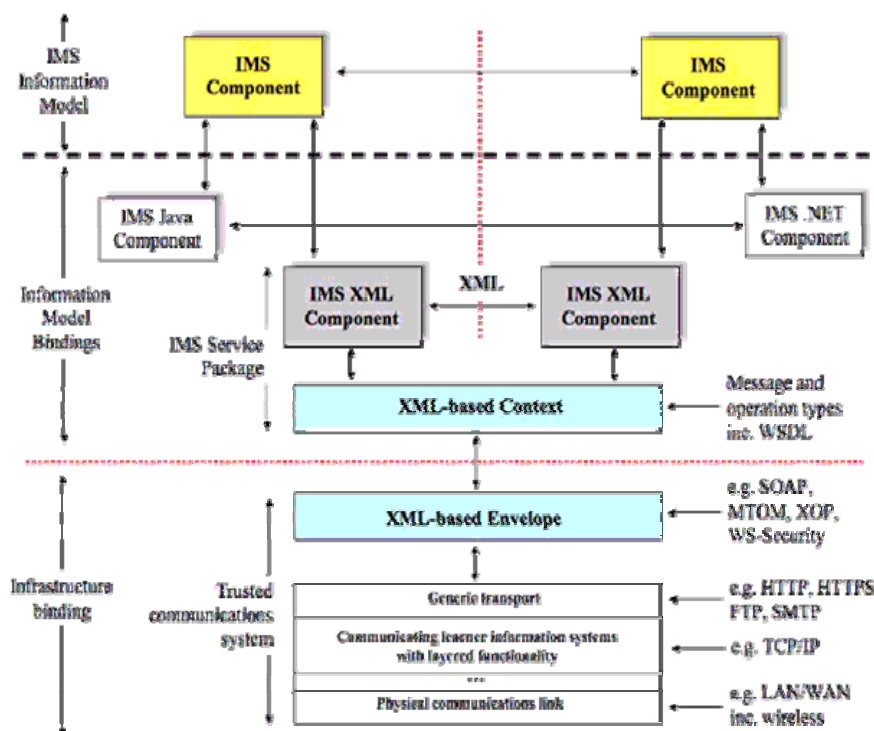


Fig. 4. La spécification IMS GWS comme mécanisme commun de transmission de messages pour des services IMS. [4]

## 6 Architecture et modélisation

La norme IMS GWS nous permet d'implémenter toute une plate forme sous la forme de web services, mais dans notre projet, nous avons développé le prototype d'un éditeur de cours conforme SCORM en IMS GWS.

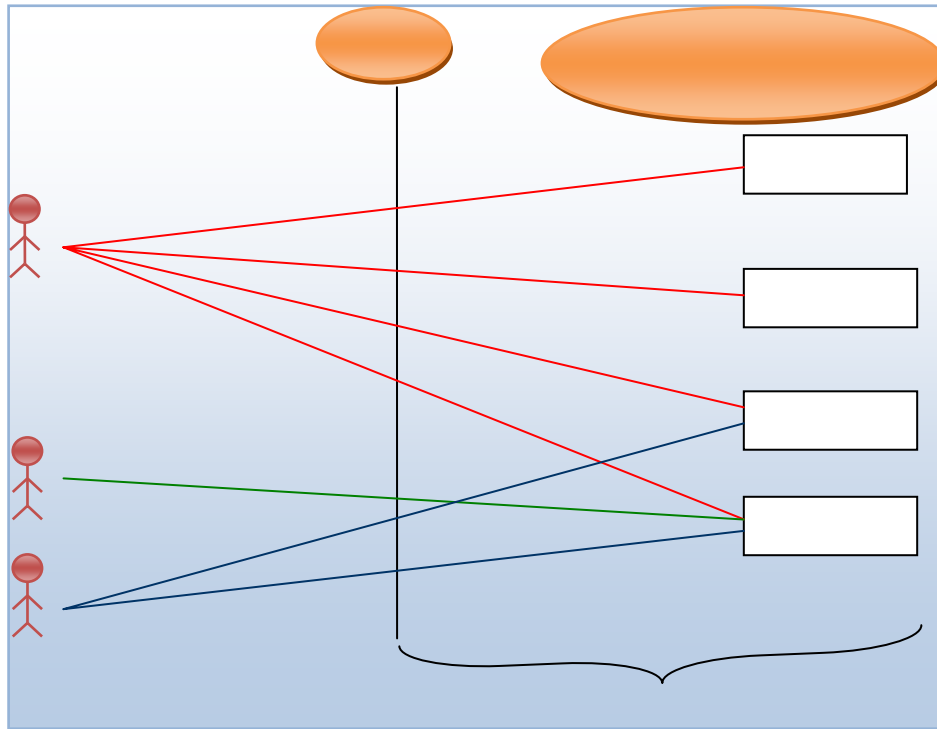


Fig. 2. Interaction entre les différents acteurs avec les Services Web via le LMS.

La modélisation a été effectuée avec UML (Unified Modeling Language) et WAE (Web Application Extension) pour la modélisation des web services,

La conception d'application Web se distingue de la conception d'autre système par deux activités majeures : la répartition des objets sur le client ou le serveur et la définition de l'interface utilisateur sous forme de pages web. L'extension d'UML pour les applications web (le WAE) nous fournit une notation pour exprimer les composants de technologie web du système dans le modèle. [10] [11]

## 7 Implémentation

Nous avons choisi d'implémenter notre application avec le langage JAVA vu que ce dernier est un langage simple, intuitif, orienté objet, performant et dynamique.

Pour la plateforme web services, nous avons utilisé l'architecture J2EE [5] qui est basée sur le langage Java qui permet le déploiement de composants sur diverses plates-formes indépendamment de leur langage de programmation.

Dans notre implémentation on utilise le mixage entre les deux produits d'Apache : **TomCat** et **Axis** pour donner à TomCat la possibilité de déployer des classes JAVA comme des Web Services. [6]

Pour implémenter notre éditeur nous avons eu besoin de quatre services web :

### 1. Le service web Consultation:

Il permet aux utilisateurs de consulter l'ensemble des ressources pédagogiques existantes dans la base de données de l'éditeur et aussi de faire des recherches selon plusieurs critères (par titre, par auteur...etc.).



Fig. 3. Exemple : Service Web consultation

## 2. Le service web Edition:

Il permet aux concepteurs de faire des mises à jour telles que l'ajout, la modification et la suppression des ressources



Fig. 4. Exemple : Service Web consultation

## 3. Le service web Exerciceur:

Dans la norme SCORM les questions sont aussi des ressources, leur modification et suppression se font de la même manière que dans le service web Edition. Par contre leur ajout diffère d'où la nécessité d'utiliser un service Exerciceur qui permet aux concepteurs d'ajouter des questions et des exercices, et d'afficher leur liste.



Fig. 5. Exemple : Service Web Exerciseur

#### 4. Le service web Exportation:

Il permet de créer le package à partir d'un ensemble de ressources, de le zipper et de l'exporter pour pouvoir le réutiliser ailleurs (dans multiple).



Fig. 6. Exemple : Service Web Exportation

## 8 Conclusion

Notre étude a montré qu'un système d'enseignement à distance peut être décomposé en plusieurs activités réalisées comme des applications autonomes sous la forme de services web.

Dans notre projet, nous avons travaillé sur la partie éditeur (Contenu). Nous avons défini les composants et les fonctionnalités de base d'un éditeur de cours, en se basant sur le standard SCORM pour la structuration, l'organisation et l'interopérabilité des contenus pédagogiques et sur la normalisation IMS GWS pour l'interopérabilité des systèmes (plates forme).

Initialement la norme IMS GWS a été conçue pour uniformiser les implémentations des différentes spécifications d'IMS qu'on a vu précédemment dans le paragraphe 4. Après l'étude de cette norme, nous avons pensé à implémenter des services web (compatibles à la spécification IMS GWS) qui pouvaient être invoqués par de multiples plates formes. Ce qui donne une flexibilité, Interopérabilité et réutilisabilité des différentes fonctionnalités de l'éditeur et c'est le point fort de sa performance.

Notre approche s'appuie sur des principes simples en utilisant des concepts standards et outils *open source*.

Nous avons représenté un exemple d'implémentation des services web dans un système *e-Learning* pour mettre en place un cours pédagogique. L'environnement que nous avons réalisé reste dans les normes des objets pédagogiques et permet aux étudiants de pratiquer des apprentissages divers. Un prototype du système est en cours de test et d'évaluation qui se fera auprès d'apprenants et de concepteurs.

## References

1. A El Maouhab, F Medjek et O Benabderrahim, "Les learning management systems (LMS) compatibles SCORM (Sharable Content Reference Model)", Projet fin d'étude CERIST, USTHB, 2004,
2. Jérôme Daniel, Vuibert, " Services web Technique, démarche et outils", 2003.
3. e-learning et services web  
[www.revue-eti.net](http://www.revue-eti.net)
4. IMS Content Packaging Specification, IMS GWS Primer: Introduction à la spécification IMS GWS, IMS GWS Security Profile: Prolongation du profil de base pour supporter différentes architectures de sécurité, IMS GWS WSDL: Langage permettant d'exposer dans un format XML la signature d'un service web accessible sur internet.  
<http://www.imsglobal.org>
5. J2EE, Java 2 Enterprise Edition  
<http://ashita-studio.com/articles/j2ee/>
6. Apache, TomCat et Axis  
<http://www.apache.org/>
7. e-learning as a web service  
<http://www.dbs.cs.uni-duesseldorf.de/gvd2004/papers/WesterkampPeter.pdf>
8. Web Service Architecture for e-learning <http://grids.ucs.indiana.edu/ptliupages/publications/E388NH.pdf>
9. Project Advanced Distributed Learning, le modèle de référence SCORM (Sharable Content Object Reference Model)  
<http://www.adlnet.org>
10. Conallen Jim, "Conception des applications web avec UML", Edition Eyrolles Paris, 2004
11. Web Application Extension for UML  
<http://www.rationalrose.com/stereotypes/waasetup.htm>

## *Ontologies engineering & reverse engineering*

---



## Une extension de l'ontologie de domaine pour supporter la multitude de contextes lors de la spécification des besoins

Achraf MTIBAA<sup>1</sup>, Wassim JAZIRI<sup>2</sup>, Faïez GARGOURI<sup>3</sup>

Laboratoire MIRACL  
Institut supérieur d'informatique et du multimédia de Sfax - Tunisie  
BP 242 - 3021 Sfax – TUNISIE

<sup>1</sup>achraf.mtibaa@issatgb.rnu.tn

<sup>2</sup>wassim.jaziri@isimsf.rnu.tn

<sup>3</sup>faiez.gargouri@fsegs.rnu.tn

**Résumé.** L'objectif général de ces travaux est de s'orienter vers les mécanismes de multi-représentation rencontrés lors de l'étape de spécification des besoins utilisateurs pour les adapter aux différents contextes. Les besoins utilisateurs sont généralement définis selon différentes représentations et accordés à plusieurs contextes. Par conséquent, les responsables de la réalisation de ces besoins se trouvent souvent confrontés aux problèmes sémantiques. Pour remédier à ces problèmes et pour assister les futurs utilisateurs dans l'expression de leurs besoins, nous orientons notre réflexion vers l'utilisation des ontologies connues par leurs apports incontestables au niveau sémantique. Cependant, une ontologie devrait fournir des définitions et des structures de données contextuelles pour représenter la diversité des perceptions et des réflexions. Ce volet n'est pas traité dans la plupart des travaux actuels sur les ontologies. L'objectif de ce papier est de proposer une couche ontologique contextuelle pour supporter la multitude de contextes lors de la spécification des besoins. Cette couche ajoutée à une ontologie de domaine construit une ontologie de multi-représentation. Cette dernière redéfinit un concept par un ensemble variable de propriétés selon le contexte dans lequel le concept de besoin est utilisé.

**Mots-clés :** Ontologie, Multi-représentation, Spécification des besoins, Multi-contexte, Couche ontologique contextuelle.

### 1 Introduction

L'étape de spécification des besoins (SB) représente une étape importante intervenant pendant la phase d'analyse des besoins du cycle de vie d'un système. Cette étape est, selon Tudor [1], un contrat entre les futurs utilisateurs et les concepteurs. Elle concerne les caractéristiques attendues (exigences fonctionnelles et non fonctionnelles) par l'utilisateur du système.

La variation de l'environnement et du contexte d'usage du système d'information (SI) peut entraîner le changement des besoins des utilisateurs, voire même leurs points de vue et attitudes vis-à-vis d'une situation (d'une prise de décision) donnée. La réaction d'un utilisateur et la décision qui la déclenche peuvent donc être influencées par différents aspects, en l'occurrence ses compétences et caractéristiques personnelles (profil) et la situation dans laquelle il se trouve (le contexte d'usage au moment de la prise de décision). Le profil lui-même peut être influencé par le contexte de l'utilisateur. Cette multitude de contextes entraîne une expression de besoins différents, voire contradictoires. Le concepteur, dans ce cas, doit faire face à divers problèmes liés notamment à l'incohérence, à l'ambiguïté sémantique et à la difficulté de modélisation des besoins.

D'un autre côté, les ontologies sont devenues une importante orientation de recherche, notamment dans l'ingénierie des connaissances, le Web sémantique, la médecine, etc. Une ontologie peut être décrite comme une spécification explicite d'une conceptualisation [2]. Cette conceptualisation permet d'identifier, par un processus d'abstraction, les concepts ou les termes essentiels d'un domaine. La spécification rend explicite la sémantique ou le sens associé aux termes [3].

La question qu'on se pose, dans notre travail, est donc comment adapter l'ontologie pour prendre en compte la variation de contextes d'utilisateurs dans un domaine donné pour spécifier leurs besoins ? Les entités présentées dans la plupart des ontologies possèdent une seule représentation modélisée dans un seul contexte. En réalité, les différentes représentations possibles d'une même entité du monde réel sont

causées par plusieurs facteurs. En particulier, on peut citer : les divers mécanismes d'abstraction considérés par les utilisateurs, la diversité des points de vue, du niveau de détail et d'intérêt de l'utilisateur, etc.

Selon Cullot [4], l'information peut être une entité sujette à différentes représentations selon le contexte dans lequel elle est considérée. Beaucoup de recherches montrent que la sémantique est dépendante du contexte dans lequel les concepts sont employés [5]. L'ontologie devrait alors fournir des structures de données contextuelles pour représenter la diversité des perceptions et des centres d'intérêt. Elle joue un rôle essentiel dans l'interopérabilité sémantique des systèmes d'information. Cette interopérabilité sémantique n'est pas facile à réaliser car la connaissance peut être décrite par plusieurs termes en utilisant des suppositions et des structures de données différentes. Nous estimons que les ontologies de multi-représentation peuvent réaliser cette intégration. En effet, ces ontologies offrent l'aspect dynamique nécessaire pour supporter les différents contextes des utilisateurs.

Nous proposons, dans cet article, une extension d'une ontologie de domaine par une couche ontologique contextuelle pour supporter la multitude de contextes lors de la SB. La suite de cet article est organisée comme suit. La section 2 précise davantage la problématique abordée dans nos travaux de recherche. La section 3 présente l'extension de l'ontologie de domaine par la couche ontologique contextuelle. Le méta-modèle de l'ontologie de multi-représentation est présenté dans la section 4. La section 5 expose un cas pratique du domaine du e-learning. Nous terminons par la conclusion et les perspectives de ce travail.

## **2. Problématique**

Il n'existe pas une seule manière pour voir (et par suite représenter) une information en raison de la multitude des perceptions et des représentations qui pouvant être accordées au contexte d'usage. Chaque point de vue implique une représentation différente de la réalité. Dans le cas de la SB, selon différents utilisateurs, la notion de multi-représentation prend toute sa dimension. En effet, chaque utilisateur, admettant un contexte particulier, peut spécifier ses besoins différemment. Cette multitude d'interprétations possibles d'un même concept de besoin s'avère gênante, en particulier, dans le cas d'un travail de conception collaboratif. Lors de cette étape, plusieurs constats peuvent être faits. Un même concept peut être présenté selon plusieurs formalismes, sous une forme qui varie en fonction du contexte et de la représentation. Il peut être interprété différemment selon le contexte dans lequel il est spécifié. Une fois défini, un concept peut être associé à un nombre illimité de représentations.

Cette multitude de contextes, utilisant des formalismes différents, amène les réalisateurs du système à des problèmes possibles de redondance et d'incohérence. Ces conflits peuvent compliquer considérablement la réalisation du système. Ces derniers ont motivé la recherche de solutions pour les résoudre et pour assister les futurs utilisateurs à spécifier leurs besoins multi-contexte.

Bien que les ontologies soient considérées comme un bon moyen de résoudre ce genre de problèmes, la plupart d'entre elles (actuellement) ne représentent pas un moyen efficace pour prendre en compte la notion de contexte. Un contexte peut être vu en tant que divers critères tels que le paradigme d'abstraction, le degré de granularité, l'intérêt des communautés d'utilisateurs ou la perception du réalisateur d'ontologie. Le même domaine peut avoir plus qu'une ontologie, où chacune d'elles est décrite dans un contexte particulier. Ces ontologies sont appelées, selon Benslimane [6], une ontologie de mono-représentation. Ainsi, les concepts dans ce type d'ontologie sont définis avec une seule représentation. La principale motivation des travaux de cet auteur était de décrire une ontologie considérée selon plusieurs contextes en même temps. Une telle ontologie est appelée une ontologie de multi-représentation. Cette dernière est une ontologie qui caractérise un concept ontologique par un ensemble variable de propriétés ou d'attributs dans plusieurs contextes. Ainsi, dans une ontologie de multi-représentation, un concept est défini une seule fois mais avec plusieurs représentations. Dans notre travail, le contexte permet de choisir un sous-ensemble de l'ontologie. Ce sous-ensemble représente une interprétation des concepts d'une ontologie dans un contexte précis. Les différents composants (concepts avec leurs propriétés et leurs différentes relations) interprétés dans différents contextes associés à une ontologie de domaine (comme support de départ) représentent une ontologie de multi-représentation. Nous procédons à une modélisation multi-contextuelle des besoins. Afin de remédier à ces problèmes et assister les futurs utilisateurs dans l'expression de leurs besoins, nous combinons l'ontologie et la multitude de contextes.



Dans ce travail, nous faisons abstraction à l'étape de construction de l'ontologie, traitée par nombre de travaux. Nous focalisons notre travail sur la réutilisation d'une ontologie de domaine en ajoutant une couche ontologique contextuelle pour supporter la notion de contexte.

### 3 Couche ontologique contextuelle comme extension de l'ontologie de domaine

Nous focalisons notre travail sur l'utilisation d'une ontologie de domaine en ajoutant une couche ontologique contextuelle pour supporter la notion de contexte. Dans notre modèle, nous admettons une ontologie de domaine construite par un prototype standard connu (par exemple *protégé2000*, *PLIB Editor*, etc.). Notre nouvelle ontologie de domaine étale les spécificités des ontologies construites pour avoir une ontologie dédiée à la SB. L'ontologie de domaine est composée des termes couramment utilisés pour spécifier les besoins d'un domaine donné, les propriétés de ces termes et les relations qui les relient. Pour supporter les nouveaux besoins de multitude de contexte et de la multitude de formalisme, une couche ontologique contextuelle sera ajoutée à l'ontologie de domaine originale. Cette couche contient plusieurs interprétations possibles, sachant que chacune d'entre elles est relative à un contexte donné. De plus, chaque terme de besoin est vu différemment selon le contexte dans lequel il est considéré. L'ontologie de domaine étendue par la couche ontologique contextuelle représente notre ontologie de multi-représentation. La définition d'un concept de besoin selon plusieurs contextes dans une même ontologie représente le dynamisme sémantique de notre ontologie de multi-représentation. Ce dynamisme assuré par de telles ontologies présente le point fort de ce type d'ontologies. Nous avons présenté, dans [7], l'architecture fonctionnelle de notre environnement. Dans la figure 1, nous présentons le modèle de l'ontologie de multi-représentation avec la combinaison de ses deux couches (ontologie de domaine + couche ontologique contextuelle).

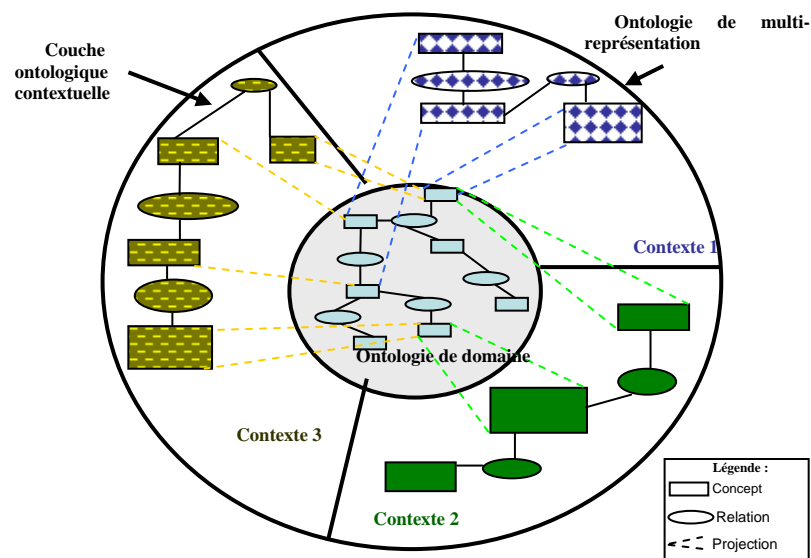


Fig. 1. Ontologie de multi-représentation

L'ontologie noyau (de domaine) est composée par des concepts représentés sous forme de rectangle et par des relations représentées par des ellipses. Les concepts de l'ontologie noyau seront projetés dans des interprétations de la couche supérieure (couche ontologique contextuelle). Dans cette couche, les concepts sont aussi reliés entre eux par des relations. Chaque interprétation est désignée par une couleur différente des autres. Chaque concept de besoin peut avoir une ou plusieurs interprétations selon le contexte considéré. Chaque concept de besoin peut admettre des relations avec les autres concepts de besoin du même contexte. Dans la section suivante, nous présentons le méta-modèle de notre ontologie de multi-représentation dédiée à la SB.

#### 4 Méta-modèle de l'ontologie de multi-représentation

Dans la figure 2, nous présentons le méta-modèle de notre ontologie de multi-représentation en suivant les annotation d'*UML* [8].

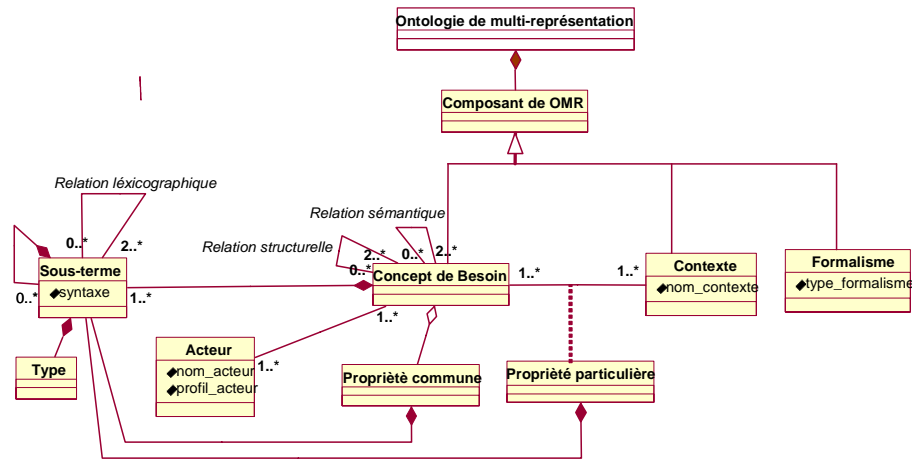


Fig. 2. Méta-modèle d'une Ontologie de multi-représentation dédiée à la SB multi-contexte

Les composants d'une ontologie de multi-représentation sont les concepts de besoin, le contexte et le formalisme. Le contexte est défini par son nom et le formalisme défini par son type (informel, semi-formel, formel). Pour un concept de besoin, il peut avoir des relations structurelles avec un autre concept de besoin. Ce type de relation est dépendant du formalisme utilisé. Dans le cas d'un formalisme semi-formel comme un diagramme des cas d'utilisation, nous avons les relations structurelles *extend*, *include* ou *est-un*. Pour le formalisme textuel, nous pouvons avoir une seule relation structurelle qui peut exister de type *est-un*. Un concept de besoin peut avoir une relation sémantique avec un autre concept de besoin. Cette relation sémantique peut être une relation de : équivalence, partie-de, sorte-de, etc. Le concept de besoin possède des propriétés communes pour tous les contextes et des propriétés particulières pour un contexte déterminé. Nous avons présenté, pour cela, une classe d'association *Propriété particulière* entre la classe *Contexte* et la classe *Concept de besoin*. Un concept de besoin est défini par un acteur. Une propriété commune ou une propriété particulière ou un concept de besoin est composée d'un ensemble de sous-terme. Ce dernier peut être un mot simple, un mot composé, un mot vide, etc. Il peut être lui-même composé d'un ensemble de sous-terme (comme par exemple *mise en œuvre*). Un sous terme peut avoir des relations lexicographiques de type : *synonymie*, *homonymie*, *polysémie*, *antonymie*, *hyperonymie*, *hyponymie*, *méronymie*, etc. Dans la section suivante, nous présentons un petit cas pratique de la multitude de contexte lors de la SB pour la conception d'un système collaboratif d'établissement d'une plate-forme de e-learning entre trois pays.

#### 5 Cas pratique : domaine du e-learning

Dans cette section, nous présentons brièvement le domaine du e-learning puis nous exposons certaines SB selon des formalismes et des contextes différents. Le domaine de notre cas pratique est l'enseignement à distance (ou *e-learnin*). Nous admettons que les principaux acteurs de notre domaine sont : *Etudiant*, *Tuteur* et *Administrateur*. Nous admettons aussi qu'il existe trois contextes dont chacun est lié à un emplacement différent (pays pour notre cas). Selon l'emplacement et l'environnement, le contexte de l'acteur peut changer du fait que les services administratifs et académiques ainsi que leur organisation peuvent varier d'une académie à une autre et d'un pays à un autre. Nous considérons trois contextes pour trois rectorats académiques différents qui désirent établir une plate-forme du e-learning dans leurs modes d'enseignement : *Rectorat de Sfax en Tunisie (RST)*, *Rectorat de l'académie de Toulouse en France*

(RATF) et Rectorat de l'université Laval en Canada (RULC). Concernant les techniques de SB, nous nous limitons à deux formalismes : informel (*SB Textuelle* : ce type de SB peut être extrait à partir des interviews ou des cahiers de charge) et semi-formel (*Diagramme des Cas d'Utilisation (DCU)* [8]).

La figure 3 présente une SB du *Tuteur* dont le contexte est de *RST* et selon le formalisme semi-formel *DCU*. Selon un autre contexte comme le cas du *RATF*, le *Tuteur* peut spécifier ses besoins différemment selon son contexte environnant. La figure 4 présente son SB selon le même formalisme *DCU*. Dans la figure 5, nous exposons la spécification textuelle exprimée par le même acteur *Tuteur* mais selon le contexte du *RULC*. Nous indiquons dans cette SB, l'acteur, le contexte, les termes de besoin (TB) et leurs opérations (OP). Ces opérations permettent de décrire les besoins des utilisateurs. Chaque terme de besoin est décrit par un ensemble d'opérations pour le cas du formalisme *Textuel* et décrit par un ensemble d'actions pour le cas du formalisme *DCU* (ne sont pas données dans cet article).

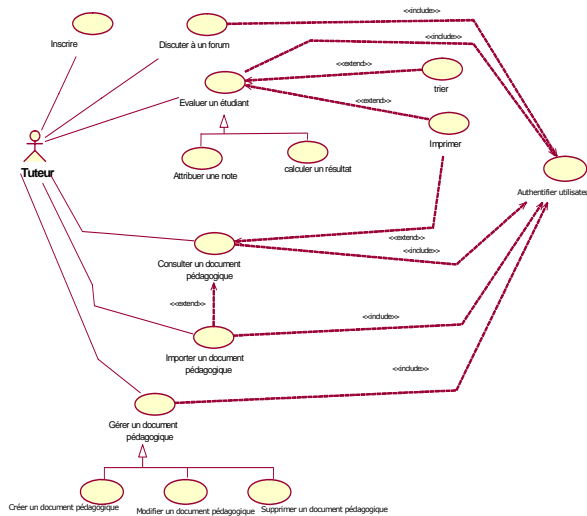


Fig. 3. SB de l'acteur Tuteur selon le contexte RST avec le formalisme DCU

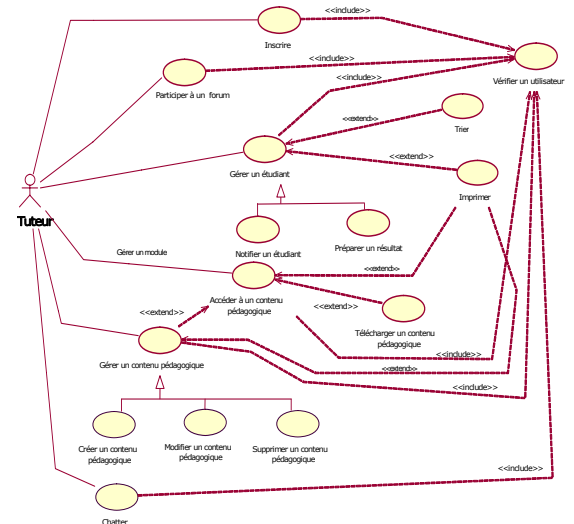


Fig. 4. SB de l'acteur Tuteur selon le contexte RATF avec le formalisme DCU

Spécification des besoins textuelle	
Acteur : Tuteur	
Contexte : Rectorat de l'université Laval au Canada	
Terme de Besoin (TB) 1 : Inscrire	Opération (OP) 1 : Remplir un formulaire. OP 2 : Choisir d'une catégorie
TB 2 : Authentifier	OP1 : Saisir un login et un mot de passe
TB 3 : Gérer un examen	OP1 : Saisir un examen OP2 : Enregistrer un examen OP3 : Elaborer une degré de difficulté OP4 : Corriger un examen OP5 : Attribuer une note
TB 4 : Communiquer via un forum	OP 1 : Choisir un thème désiré OP2 : Participer à une vidéoconférence
TB 5 : Gérer un document électronique	OP1 : Demander de gérer un document électronique OP2 : Consulter un document électronique OP3 : Attacher un document électronique OP4 : Modifier un document électronique OP5 : Supprimer un document électronique OP6 : Imprimer un document électronique
TB 6 : Gérer un résultat	OP1 : Saisir un coefficient d'une matière OP2 : Calculer une moyenne OP3 : Afficher un résultat OP4 : Imprimer un résultat

Fig. 5. SB de l'acteur Tuteur selon le contexte RULC avec formalisme textuel

Dans les trois SB, nous remarquons la diversité des représentations des besoins du même acteur en admettant les contextes des trois rectorats. Nous pouvons avoir pour chaque acteur deux formalismes et trois contextes. Nous avons, en totalité, dix-huit SB<sup>1</sup>. Avec cette énorme hétérogénéité et pour des raisons de simplification, nous avons présenté seulement trois SB. Pour réaliser une plate-forme du e-learning en satisfaisant tous les besoins de tous types d'utilisateur, nous proposons l'utilisation d'une ontologie de domaine associée à une couche ontologique contextuelle. Cette ontologie (de multi-représentation) est composée par les concepts de besoins selon les trois contextes. Elle assistera les futurs utilisateurs pour spécifier leurs besoins. Nous avons commencé par une étude comparative entre les différents besoins des différentes SB. Nous avons pu vite conclure l'énorme problème auquel sont confrontés les concepteurs et les réalisateurs d'une telle plate-forme. La mise en application d'une ontologie du domaine du e-learning via un éditeur standard avec une couche contextuelle projetant les concepts de besoin dans les différents contextes est une tâche en cours de réalisation.

## 6 Conclusion

Nous avons proposé, dans cet article, une couche ontologique contextuelle associée à une ontologie de domaine. Ces deux composants représentent notre ontologie de multi-représentation dédiée à la SB multi-contexte. Nous avons présenté le méta-modèle de cette ontologie. Dans un futur travail, nous désirons formaliser le méta-modèle et développer un environnement qui intègre notre ontologie pour assister les utilisateurs dans l'expression de leurs besoins multi-contexte.

## Références

1. Tudor, H. : Guide d'ingénierie des exigences logicielles avec UML. Projet: SPINOV, Guide pratique, (2005)
2. Gruber, T. : Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human and computer studies*, Vol 43, n°5, pp. 907-928, (1995)
3. Leclercq, E. : Interopérabilité sémantique des systèmes d'information géographique : une approche basée sur la médiation de contexte. Thèse de Doctorat, Université de Bourgogne, Dijon, (2000)
4. Cullot, N., Parent, C., Spaccapietra, S., Vangenot, C. : Des SIG aux ontologies géographiques. *Revue Internationale de Géomatique*, Volume 13 - No 3/2003 - Les SIG sur le web, pp. 285 – 306, (2003)
5. Pierra, G. : Context Representation in Domain Ontologies and Its Use for Semantic Integration of Data, S. Spaccapietra (Ed.) : *Journal on Data Semantics X*, LNCS 4900, pp. 174–211, Springer-Verlag Berlin Heidelberg, (2008)
6. Benslimane, D., Vanguenot, C., ROUSSEY, C. : Multirepresentation in ontologies. *Advances in Databases and Information Systems*, Lecture notes in Computer Science, vol 2798, pages 4-15. ISSN: 0302-9743, (2003)
7. Mtibaa, A., Gargouri, F. : A multi-representation ontology for the specification of multi-context requirements, In the international conference on Signal Image Technology & Internet based Systems, pp 287-296, (2006)
8. Muller, P.A., Gaertner, N. : Modélisation objet avec UML, Edition Eyrolles, deuxième Edition, (2000-2001)

---

<sup>1</sup> 3 Acteurs \* 3 Contextes \* 2 Formalismes  $\Rightarrow$  18 SB

# Classification de Techniques d'Appariement d'Ontologies pour le Web Sémantique

Samira Oulefki<sup>1</sup>, Karima Akli-Astouati<sup>2</sup>

Département d'informatique  
Université des sciences et de la technologie "Houari Boumédienne"  
Laboratoire de Recherche en Intelligence Artificielle  
Boîte postale 32, El Alia Bab Ezzouar, Alger, Algérie

<sup>1</sup> samira.oulefki@yahoo.fr

<sup>2</sup> akli\_k@yahoo.com

**Résumé.** L'appariement d'ontologies consiste à trouver les relations entre les entités exprimées dans des ontologies différentes. Ces relations sont découvertes par le biais de mesures de similarité. Dans la littérature, plusieurs techniques pour mesurer la similarité ont été utilisées et de nombreuses classifications ont été proposées. Ces travaux abordent le problème d'appariement à partir de plusieurs domaines tels que l'intelligence artificielle, les systèmes d'information, les bases de données et le web sémantique. Dans le cadre du web sémantique, nous avons proposé une classification de techniques d'appariement, sous forme d'un modèle en couche, qui s'appuie sur les différentes caractéristiques supportées par les langages du web sémantique.

**Keywords:** techniques d'appariement, mesures de similarité, modèle en couche, correspondance sémantique, ontologies, web sémantique

## 1 Introduction

Le web qu'on essaye de construire aujourd'hui, se veut un web ayant un sens, de façon que les ressources soient accessibles à la fois par des humains aussi bien que par des agents logiciels. Les ontologies permettant une représentation explicite du sens ainsi que les métadonnées sont au centre des travaux de ce nouveau web que l'on qualifie de web sémantique. La diversité des formats de représentation de la connaissance et des modélisations pour un même domaine conduit à l'apparition de problèmes d'hétérogénéité entre différentes ontologies.

Cette hétérogénéité posera des problèmes dans l'échange, le traitement, l'intégration, et la recherche d'information. Afin de cohabiter avec cette hétérogénéité, des correspondances sémantiques entre les éléments des ontologies doivent être spécifiées. Ces correspondances constituent une "colle" qui maintient les ontologies ensemble afin de garantir leur interopérabilité. Le processus permettant de découvrir ces correspondances est appelé appariement.

Le but de l'appariement des ontologies est donc de trouver les relations entre les entités exprimées dans des ontologies différentes [5], [1], [4]. Ces relations sont, généralement, découvertes par le biais de mesures de similarité. Dans la littérature, plusieurs techniques pour mesurer la similarité entre les entités des ontologies ont été développées. Ces techniques sont dites de base car elles sont appliquées sur un aspect particulier des entités, tel que le nom, les attributs et les relations avec d'autres entités [1], [4]. Sur chaque aspect, les caractéristiques d'une entité sont comparées avec les caractéristiques correspondantes d'une autre entité, cela retourne une valeur de similarité [1].

Cependant, les langages de représentation des ontologies n'ont pas tous la même puissance d'expressivité. Par exemple, contrairement au langage OWL, le langage RDF ne permet pas de représenter les hiérarchies de subsumption. Par conséquent, le choix des techniques appropriées dépendra du langage de représentation des ontologies utilisées. La question est donc de savoir comment guider le choix d'une technique pour un aspect donné ?

Afin de répondre à cette question, cet article présente une classification des techniques de base qui sert à guider l'utilisateur dans le choix de la technique appropriée pour comparer l'ensemble des caractéristiques supporté par le langage de représentation des ontologies utilisé.

Pour cela, l'article est organisé comme suit. La section 2 présente les travaux proches traitant de classification des techniques d'appariement de base. Dans la section 3, l'hypothèse sur laquelle nous nous appuyons dans notre classification est donnée. La section 4 est consacrée à la présentation de notre approche de classification. Une conclusion terminera cet article.

## 2 Travaux proches

Dans la littérature, plusieurs classifications des techniques de base d'appariement ont été proposées [12], [4], [2]. On se limitera à celles qui sont relatives au web sémantique.

Les auteurs de [4] se basent principalement sur la classification de [12] en la complétant par l'introduction de nouvelles classes. Ils proposent quatre directives sur lesquelles cette classification se fonde, à savoir : L'exhaustivité, la disjointure, l'homogénéité et la saturation.

De plus, dans la couche des techniques de base, les auteurs distinguent les techniques utilisant des ressources externes de celles utilisant des ontologies de support. Toutefois, ces deux techniques sont liées par la relation de subsumption. Par conséquent, la directive de disjointure de classes est rompue. On retrouve également, un risque de rupture concernant la directive de saturation. En effet, l'appariement de règles, dans le cas du langage ontologique SWRL, ne peut trouver place dans cette classification et exigera donc l'introduction d'une nouvelle classe de techniques spécifiques qui s'adaptent à cette catégorie.

En ce qui concerne la classification [2], elle comprend deux catégories principales, à savoir : les techniques linguistiques et les techniques contextuelles. Une première critique peut être faite sur la disjointure non vérifiée entre les définitions des deux classes. En effet, les techniques linguistiques prennent en compte les contextes de concepts en termes de propriétés et de relations sémantiques, ce qui est couvert également par les techniques contextuelles puisqu'elles utilisent des métriques basées sur les structures graphiques ou taxonomiques. Toutefois, cette classification ne couvre pas toutes les techniques de base existantes. Par exemple, pour déduire la similarité entre les instances, elle ne considère que les techniques basées sur un raisonnement probabiliste malgré l'existence d'autres approches telles que celles basées sur les techniques statistiques [3] ou de l'analyse de données.

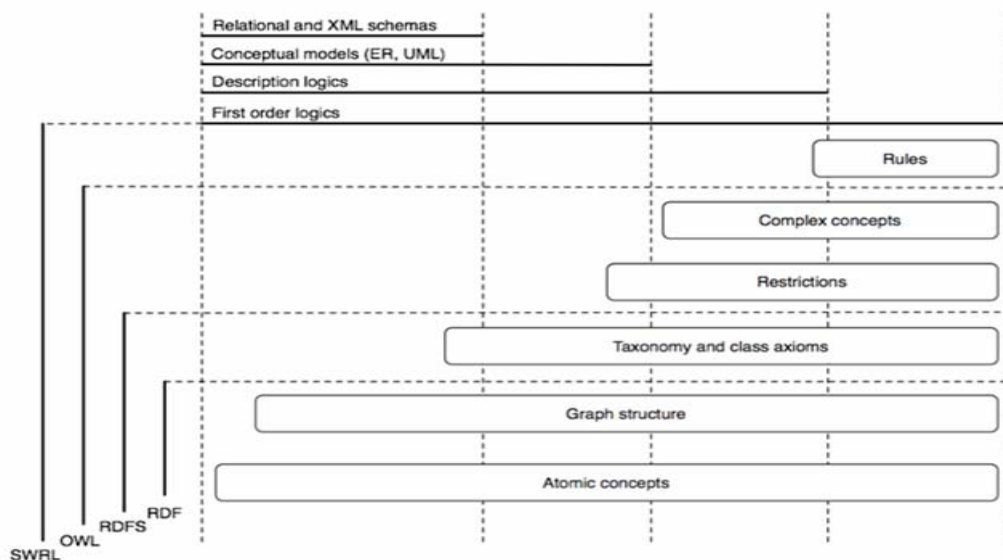


Fig. 1. Représentation graphique de la complexité sémantique [2]

*Les concepts atomiques.* Au premier niveau, les entités sont dénotées seulement par un nom, une étiquette et un commentaire. Les caractéristiques telles que les attributs ou les relations ne sont pas considérées. Tous les langages ontologiques fournissent un ensemble de symboles qui peuvent être adoptés pour représenter les concepts atomiques.

*Les structures de graphe.* Au deuxième niveau, les ontologies sont vues comme des structures de graphes où les concepts sont les nœuds et les relations sont les arcs. Dans cette vue, les concepts sont caractérisés par leur structure qui est déterminée par les arcs les reliant avec d'autres concepts ou avec les types de données simples. Les arcs sont des relations génériques entre les concepts et peuvent être représentés par des attributs ou des propriétés. Les langages ontologiques fournissent toutes les primitives permettant de représenter les graphes.

*Les classes d'axiomes et taxonomie.* À ce niveau, les ontologies sont vues comme une taxonomie de concepts. Des axiomes de classes, exprimant une relation sémantique entre les classes, peuvent être définis. Les relations typiques qui sont exprimées à ce niveau sont :

- L'équivalence qui affirme que deux classes sont composées par les mêmes instances,
- La relation is-a, appelée aussi relation de sous-classe, qui déclare qu'une classe donnée est une spécialisation d'une autre classe, c'est à dire, qu'elle contient un sous-ensemble des instances de l'autre classe.

Concernant les langages du web sémantiques, seul RDF ne supporte pas la définition de ce niveau.

*Les restrictions.* Dans ce cas, la description des ontologies peut être enrichie par des restrictions sur les propriétés. Une restriction dénote une contrainte qui est appliquée à une propriété quand cette dernière est appliquée à une classe. Fondamentalement, il y a deux types de restrictions:

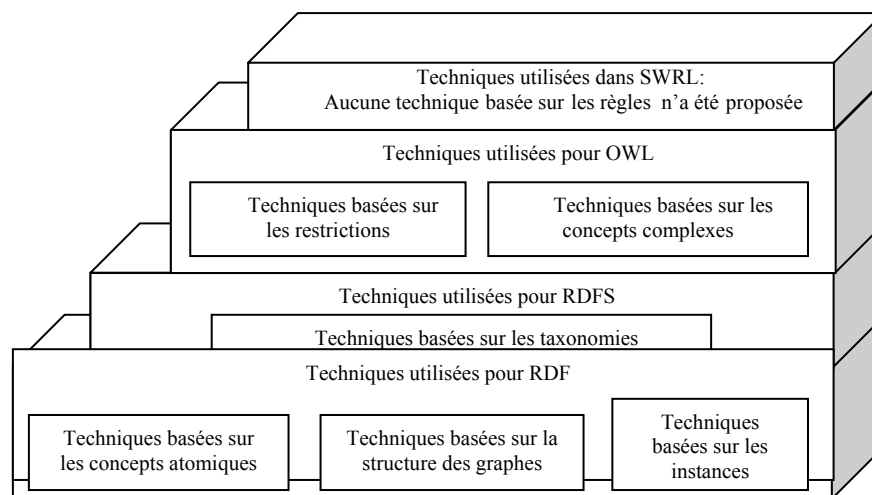
- Les restrictions de cardinalité qui dénotent le nombre minimal et maximal des valeurs de types de données et des instances pour une propriété
- Les restrictions du co-domaine expriment le fait que les classes des instances sont considérées comme des valeurs valides pour une propriété. Ce type de restrictions peut déclarer que la contrainte est valide pour tous les événements de la propriété ou qu'une propriété existe avec cette contrainte.

Dans les langages du web sémantiques, un support complet pour les restrictions est seulement disponible avec OWL.

*Les concepts complexes.* Les concepts complexes sont exprimés à l'aide de concepts atomiques relié par les opérateurs tels que l'intersection, l'union ou le complément. La définition de concepts complexes n'est possible qu'en utilisant OWL.

*Les règles.* Une règle est une expression du langage qui, étant donné un ensemble de déclarations, autorise d'inférer de nouvelles connaissances dans l'ontologie. Les règles sont seulement disponibles au niveau de SWRL.

En se basant sur ces différents niveaux de complexité sémantique, nous proposons une nouvelle classification qui est présentée par le modèle en couche suivant.



**Fig. 2.** Modèle en couche pour classifier les techniques d'appariement des ontologies

### 3 Principe de la classification proposée

Pour décrire les caractéristiques des ontologies, six (6) niveaux de complexité sémantique ont été identifiés [2]. Ils sont adressés par chaque langage du web sémantique et par conséquent par leur type respectif (Fig. 1). Chaque langage fournit des constructions pour exprimer les caractéristiques correspondant à son niveau.

#### 4.1 Techniques utilisées pour apparier des ontologies écrites en RDF

##### 4.1.1 Concept Atomique

Le nom et l'étiquette sont habituellement représentés par des chaînes de caractères courtes [1]. Tandis que, les commentaires sont représentés par des chaînes de caractères longues [1]. Les mesures de similarité proposées sont différemment conçues pour chaque catégorie :

- Les chaînes de caractères courtes sont généralement comparées en utilisant des techniques syntaxiques et/ou basées sur la structure d'une ressource externes.
  - Les techniques syntaxiques considèrent deux entités similaires si leurs noms ou leurs étiquettes sont similaires syntaxiquement [1].
  - Les techniques basées sur la structure d'une ressource externes, qui peut être une ontologie de support, opèrent en deux étapes [14]. Dans une première étape dite « d'ancrage », les termes sont apparés aux entrées d'une ressource externe. Dans la deuxième étape dite « de dérivation », les relations entre les entités sont dérivées en utilisant les relations découvertes durant l'étape d'ancrage.
- Les chaînes de caractères longues sont comparées en utilisant des techniques dites « techniques basées sur les tokens » [5], [1], [4]. Ces dernières découpent les chaînes de caractères en plusieurs morceaux appelés « tokens » [1]. L'intégration des techniques basées sur la structure d'une ressource externes dans le calcul de la similarité entre les chaînes de caractères longues n'est pas utile et est coûteuse en temps de calcul et donne comme résultat plus de bruit que d'informations utiles [1]. Les commentaires et les descriptions contiennent assez d'information descriptive pour pouvoir être comparés aux autres en employant les techniques basées sur les tokens [1], [4].

##### 4.1.2 Structure de graphe

Ces techniques exploitent des informations structurelles lorsque les entités en question sont reliées entre elles par des liens sémantiques ou syntaxiques, formant ainsi un graphe des entités [1]. Il existe deux types de relations [5], [4]: les relations méréologiques et toutes les relations impliquées :

*Structures méréologique* : L'exploitation de la structure méréologique présente la difficulté de trouver les propriétés qui portent cette structure [4]. Cependant, s'il y a lieu de détecter les relations qui supportent la structure part-of, il sera possible de calculer la similarité entre les concepts en considérant qu'ils seront plus similaires s'ils partagent des parties similaires [4].

*Les relations*: Les concepts sont aussi liés par la définition de leurs propriétés. Ces propriétés correspondent aux arcs du graphe représentant une ontologie. Par conséquent, considérer ces propriétés pour déduire la similarité entre les entités peut être intéressant [1], [4]. L'idée est que si l'on a deux entités similaires A et A', et si elles sont connectées par un même type de relation R avec deux autres entités B et B', alors on peut déduire que B et B' sont d'une façon ou d'une autre similaires. De même, si on sait que A et A' sont similaires et que B et B' sont aussi similaires, alors les relations de A-B et de A'-B' peuvent être similaires.

##### 4.1.3 Techniques utilisées pour les instances

La similarité entre deux entités, qui sont notamment des concepts, est décidée par les méthodes extensionnelles en analysant leurs extensions, c.-à-d., l'ensembles de leurs instances [1], [4].

Généralement, les méthodes extensionnelles sont divisées en deux catégories [4] : celles qui s'appliquent aux ontologies avec des ensembles d'instances communes, et celles qui travaillent sur des ensembles d'instances hétérogènes.



## 4.2 Techniques utilisées pour appairer des ontologies écrites en RDFS

En plus des techniques décrites précédemment, le processus d'appariement des ontologies écrites en RDFS peut utiliser les techniques basées sur les taxonomies. Ces dernières exploitent la relation *is-a* entre les concepts [2]. Si les concepts sont les mêmes, ils auront probablement les mêmes super-concepts, et si les super-concepts sont les mêmes, les concepts sont alors semblables. Les sous-concepts de deux classes semblables seront aussi semblables. Si les sous-concepts sont les mêmes, les concepts comparés sont semblables. Aussi, si les concepts ont des frères et sœurs semblables, ils sont aussi semblables [2].

## 4.3 Techniques utilisées dans OWL

On retrouve deux niveaux associés à OWL. Le premier est consacré aux restrictions. Tandis que le deuxième niveau soulève l'aspect de concepts complexes.

### 4.3.1 Techniques utilisées au niveau des restrictions

Ces techniques sont aussi référencées dans [1], [5] sous l'appellation « *méthodes basées sur la structure interne* » et dans [12] comme « *méthodes basées sur les contraintes* ». Afin de mesurer la similarité entre les entités, ces méthodes utilisent les informations concernant les propriétés des entités telles que le co-domaine, la cardinalité ou la multiplicité [1], [4]. L'idée est que les classes avec des propriétés associées aux mêmes contraintes ou des contraintes compatibles sont considérées semblables [4].

### 4.3.2 Techniques utilisées au niveau des concepts complexes

Afin de déduire la similarité entre deux entités, les méthodes utilisées dans ce niveau se basent sur des modèles de logique tels que la satisfiabilité propositionnelle (SAT), la SAT modale ou les logiques de description [1], [4].

## 4.4 Techniques utilisées pour appairer des ontologies écrites en SWRL (niveau des règles)

Les règles peuvent devenir aussi intéressantes pour les considérations de la similarité. Si des règles semblables entre les concepts existent, ces concepts seront considérés comme semblables [2]. Malheureusement, il n'y a pas eu de recherche suffisante et de support pratique concernant la couche des règles pour les considérations de la similarité [2].

## 5 Conclusion

Dans ce papier, nous avons présenté une classification des techniques d'appariement de base en partant des niveaux de complexité sémantique associés aux différents langages du web sémantique. Cette classification est naturelle puisqu'elle fournit un guide d'utilisation selon le langage ontologique utilisé.

Cependant, la tâche d'appariement d'ontologies est sensé trouver des rapports entre toutes les entités des ontologies qui peuvent être des classes, des relations, des instances, des axiomes et des règles [1]. Les techniques proposées jusqu'à présent n'exploitent pas le niveau de complexité sémantique correspondant aux règles. A notre connaissance, aucune recherche n'a été menée dans ce sens. Et c'est dans ce contexte que s'inscrit la suite de notre travail. De plus, les techniques de base ne suffisent pas à elles seules à assurer l'alignement des ontologies, une combinaison des résultats obtenus par d'autres techniques (de combinaison) est à envisager.

## References

1. Bach, T. L. *Construction d'un Web sémantique multi-points de vue*. Thèse de doctorat, École des Mines de Nice à Sophia Antipolis, (2006).
2. Castano, S., Ferrara, A., Hess, G. N., & Montanelli, S. *State of the Art on Ontology Coordination and Matching*. deliverable V.1.0, Bootstrapping Ontology Evolution with Multimedia Information, (2007).

3. Cox, T. F., & Cox, M. A. Multidimensional Scaling. *Chapman and Hall*, (1994).
4. Euzenat, J., & Shvaiko, P. *Ontology matching*. (2007), Springer.
5. Euzenat, J., Bach, T., Barrasa, J., Bouquet, P., Bo, J., Dieng-Kuntz, R., et al. *State of the art on ontology alignment*. Knowledge web NoE. Knowledge web NoE, (2004).
6. Hamming, R. W. Error detecting and error correcting codes. *Bell System Technical Journal* (29), (1950), 147–160.
7. Jaro, M. Probabilistic linkage of large public health data files. *Statistics in Medicine*, (1995), 14:491–498.
8. Kalfoglou, Y., & Schorlemmer, M. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18 [1], (2003), 1–31.
9. Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady akademii nauk SSSR*, 163 (4), (1965), 845-848.
10. Lin, D. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, (1998), 296-304. Madison (WI US).
11. Monge, A., & Elkan, C. An efficient domain-independent algorithm for detecting approximately duplicate database records. *SIGMOD Workshop on Data Mining and Knowledge Discovery*. Tucson (AZ US), (1997).
12. Rahm, E., & Bernstein, P. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10 (4), (2001), 334-350.
13. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy., 10, (1995), 448– 453. Montréal, Canada.
14. Safar, B., Reynaud, C., & Calvier, F.-E. Techniques d'alignement d'ontologies basees sur la structure d'une ressource complementaire. *Ieres Journees Francophones sur les Ontologies*, (2007).
15. Winkler, W. *The state of record linkage and current research problems*. Statistics of Income Division. Internal Revenue Service Publication, (1999).
16. Wu, Z., & Palmer, M. Verb semantics and lexical selection. In *Proc. 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, (1994), 133-138. Las Cruces (NM US)

## Un Index de Contenu et de Structure pour les Documents XML

Fouad DAHAK<sup>1</sup>, Mouhand BOUGHANEM<sup>2</sup>, Amar BALLA<sup>1</sup>

<sup>1</sup> Institut National d'Informatique / LCSI - EVAL  
BP 68M, Oued Smar Alger, Algérie  
{f\_dahak,a\_balla}@ini.dz

<sup>2</sup> Institut de Recherche en Informatique de Toulouse  
Professeur université Paul Sabatier  
118 route de Narbonne 31000 TOULOUSE cedex 9  
bougha@irit.fr

**Résumé.** L'indexation dans la RI (Recherche d'Informations) classique apporte peu d'outils pour le traitement des documents semi-structurés: les représentations des documents en recherche d'information ont été conçues pour des documents "plats" et homogènes. Elles ne sont pas adaptées au traitement simultané de la structure et du contenu. Plusieurs approches d'indexation semi-structurées ont fait leur apparition pour résoudre ce nouveau défi dont la plus connue est le DataGuide. Cet article présente une approche d'indexation des documents semi-structurés basée sur le DataGuide permettant de concilier contenu et structure grâce à un lien virtuel ne faisant pas partie de l'index lui-même mais construit au moment de l'indexation et de la recherche uniquement.

**Mots-clés :** DataGuide, Approches d'indexation, Données semi-structurées.

### 1. Introduction

Avec l'apparition des données semi-structurées, la nécessité de prendre en considération leur information structurelle durant le processus de recherche devient inévitable. C'est le lien entre la structure et le contenu qui pose le nouveau défi de l'indexation des données semi-structurées. La plupart des travaux de recherche actuels combinent entre des structures d'index issues de la recherche d'information classique et des bases de données pour résoudre ce nouveau problème. Les deux structures d'index les plus connues dans la recherche d'information classique sont le fichier inverse et les fichiers de signatures [11] et l'index qui a révolutionné la recherche d'information semi-structurée est, sans aucun doute, le DataGuide [4]. Cet index est décrit par ses auteurs comme étant un résumé de chemins d'étiquettes compact et construit dynamiquement sur la base de données. Ce n'est cependant, qu'un index de structure utilisé pour faciliter la formulation de requêtes et l'accélération de leur évaluation. Le DataGuide a été combiné avec d'autres structures pour permettre l'indexation du contenu textuel des documents semi-structurés. Les nouvelles versions du DataGuide ont certainement amélioré les performances de recherche mais restent impuissantes devant le classement du résultat de la recherche par ordre de pertinence vu qu'elles ne prennent pas en considération la pondération des termes. Nous montrons dans cet article que les performances de recherche peuvent être nettement améliorées et le tri du résultat peut être possible en combinant le DataGuide avec un fichier inverse étendu. Notre index présente un lien fort et implicite entre la structure (supportée par le DataGuide) et le contenu (sauvegardé dans le fichier inverse étendu) grâce à une structure intermédiaire, dite signature virtuelle, construite uniquement au moment de l'indexation et de la recherche, ne faisant ainsi, pas partie de l'index.

### 2. Etat de l'art

Dans cette section, on se focalise principalement sur les index de contenu et de structure des documents semi-structurés : Le dernier index connu basé sur le *DataGuide* est sûrement le *Content-Aware DataGuide* [9] dans lequel l'index de contenu est représenté soit avec le fichier inverse ou les fichiers de signatures est complètement intégré dans le *DataGuide*, une partie du contenu réside en mémoire centrale permettant ainsi de réduire le nombre d'accès au disque lors de la recherche. Cet index

présente de très bonnes performances en termes de taille de l'index et de rapidité de recherche. Le *Compact Tree* ou *Ctree* [7] est une structure d'arbre très compacte basée sur le *DataGuide*. Dans cet index, chaque nœud du *DataGuide* est appelé groupe et contient tous les nœuds équivalents dans la base de données partageant le même chemin d'étiquette. Le *Ctree* fournit, non seulement, un *DataGuide* mais aussi des rapports de niveau d'élément détaillés en conservant des liens parent / enfant entre les nœuds. L'IndexFabric [2] est l'index le plus proche du *Content-Aware DataGuid*. Basé sur le *DataGuide*, il fournit un lien entre les nœuds de l'index et les occurrences des mots-clés. Ce lien est explicite : les nœuds de l'index sont enrichis d'un *Trie* représentant le contenu textuel. L'IndexFabric est également équipé d'une architecture de stockage sophistiquée. Les consultations dans l'IndexFabric peuvent suivre le modèle typique pour les index à base de graphe : un simple chemin donné est lié à un ensemble d'identifiants de nœuds. L'index BUS [3] supporte la structure combinée et les requêtes de contenu avec un classement de résultat. Parmi tous les nœuds correspondants à la partie structurale de la requête, ceux où le mot-clé de la requête se produit le plus souvent sont classés le plus haut. Comme les précédents, cet index impose cependant, une identification des nœuds de la collection. La Hiérarchie de Fichier de Signature [1] combine des fichiers de signature et des *Tries*. Cet index a un filtre de contenu incorporé permettant de renoncer, durant l'appariement de la structure aux nœuds ne contenant pas les mots-clés de la requête. On remarque que le *DataGuide* a inspiré la plupart des approches pour représenter la structure des documents, avant de présenter notre approche qui s'inspire également de cet index nous présentons dans la section suivante l'index *DataGuide* original tel qu'il a été introduit dans [6].

### 3. L'indexation avec le DataGuide original

Le *DataGuide* [4] est essentiellement une représentation concise de l'arbre de la base de données dans lequel, chaque chemin d'étiquette de la base de données apparaît exactement une seule fois. La figure 1 (b) présente le *DataGuide* de l'arbre de la base de données dans la figure 1 (a). Le *DataGuide* est souvent plus petit que l'arbre de la base de données, ce qui permet de le maintenir en mémoire centrale lors de la recherche. Sans identification des nœuds de la collection, le *DataGuide* ne peut que renseigner sur l'existence d'un chemin d'étiquettes mais ne peut en aucun cas identifier de manière unique ce chemin dans la collection. Pour cela, un identifiant est attribué à chaque nœud du *DataGuide*. Cet identifiant est ensuite sauvegardé dans une table dite table des annotations (figure 1 (c)) sur le disque et pointe vers les identifiants des nœuds des documents de la collection qu'il représente.

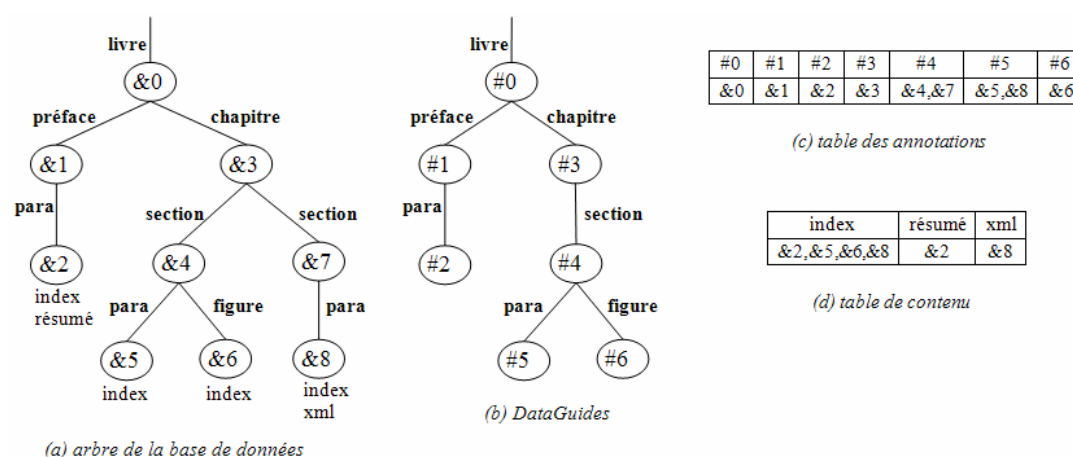


Fig. 1. Structure de l'index DataGuide

Pour prendre en considération le contenu textuel, le *DataGuide* est combiné avec un fichier inverse dans [6]. Ainsi, pour chaque mot-clé est associé un ensemble d'identifiants des nœuds de la collection où ce mot-clé apparaît et le tout est enregistré dans le disque dans une table de contenu (figure 1 (d)). L'index que nous proposons garde le principe du *DataGuide* pour la représentation de la structure et permet, en

plus, de représenter le lien père/fils entre les nœuds de niveaux différents. La section suivante présente les fondements théoriques de notre approche.

## 4. Un DataGuide annoté avec un index de contenu

L'index que nous proposons dans cet article se base sur le *DataGuide* pour l'indexation de la structure et d'une extension du fichier inverse pour la représentation du contenu en réalisant un lien virtuel entre les deux grâce à une signature virtuelle d'étiquettes. Notre index se compose donc d'un *DataGuide* annoté et d'un fichier inverse étendu.

### 4.1. Le DataGuide annoté

Le *DataGuide* annoté est un résumé de chemin d'étiquettes construit sur la base de données et dans lequel chaque chemin d'étiquette de la base de données ne figure qu'une seule fois. Chaque nœud  $N_{dg}$  du *DataGuide* annoté représente donc un ensemble de nœuds  $N_{db}$  de la base de données. Un nœud  $N_{dg}$  du *DataGuide* est annoté si et seulement si il existe au moins un nœud  $N_{db}$  de l'ensemble des nœuds de la base de données qu'il représente contenant du texte. En d'autres termes, ce ne sont que les nœuds feuilles ou mixtes de la base de données dont les représentants au niveau du *DataGuide* sont annotés. Le contenu de cette annotation représente le lien entre la structure et le contenu, il est détaillé dans les deux sections suivantes. La figure 2 (b) représente le *DataGuide* annoté de l'arbre de la base de données de la figure 2 (a) formé de deux documents  $d1$  et  $d2$ . Le *DataGuide* ne représente que la structure de la base de données, il doit donc être combiné avec une autre structure pour permettre la représentation du contenu et le plus important c'est le lien entre le *DataGuide* et cette structure. Dans notre cas, ce lien est assuré par la signature virtuelle.

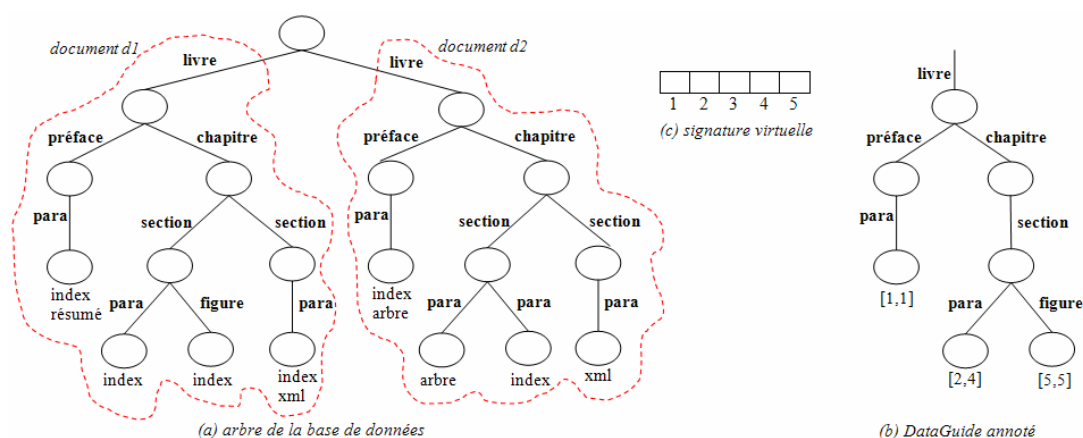


Fig. 2. Le DataGuide annoté et la signature virtuelle

### 4.2. La signature virtuelle

La signature virtuelle est une structure construite au moment de l'indexation et de la recherche uniquement. Elle ne fait pas partie de l'index et donc, pas sauvegardée sur le disque. La signature virtuelle est une suite de bits, dans laquelle, à chaque nœud annoté  $N_{dga}$  du *DataGuide* annoté correspond un ensemble de bits successifs dans la signature virtuelle (figure 2 (c)). Le nombre de bits correspondants à chaque nœud annoté  $N_{dga}$  du *DataGuide* annoté est égale au nombre maximum de fois que le chemin menant à l'étiquette représentée par  $N_{dga}$  apparaît dans les documents de la base de données. Les nœuds du *DataGuide* sont représentés avec un parcours en profondeur d'abord. L'annotation d'un nœud du *DataGuide* n'est autre que l'intervalle de type  $[minpos, maxpos]$  représentant les positions des bits de la signature virtuelle correspondant à ce nœud où  $minpos$  est la position du premier bit et  $maxpos$  la position

du dernier bit dans la signature représentant le nœud en question. En considérant les documents séparément et en suivant le parcours en profondeur d'abord, à chaque nœud  $n_i$  d'un document  $d_j$  accessible par le chemin d'étiquettes  $p$  correspond un bit dans la signature virtuelle. Cette technique nous permet d'identifier de façon unique chaque nœud des documents de la collection. La longueur de l'intervalle calculé par  $l = \maxpos - \minpos + 1$  représente le maximum de fois que le chemin menant au nœud dont l'annotation est cet intervalle apparaît dans les documents de la collection. La taille de la signature est donc égale à la somme des longueurs des intervalles dans les nœuds annotés du *DataGuide* annoté. La formule suivante permet de calculer la taille  $length(s)$  de la signature  $s$ .

$$length(s) = \sum_{i=1}^{ndg} l_i = \sum_{i=1}^{ndg} \max_{j=1}^{ndb} (Nb(c_i, d_j))$$

Où  $ndg$  c'est le nombre de nœuds annotés dans le *DataGuide*,  $ndb$  c'est le nombre de nœuds ayant un contenu textuel,  $ndb$  est le nombre de documents dans la collection,  $Nb()$  renvoie le nombre d'apparition du chemin d'étiquette  $c$  dans le document  $d$  et la  $max()$  renvoie le maximum des  $Nb$  pour tous les documents.

#### 4.3. Le fichier inverse étendu

Pour l'indexation du contenu textuel des documents de la collection on construit pour chaque mot-clé, pour chaque document une signature virtuelle en mettant à 1 les bits représentant les nœuds du document où ce mot-clé apparaît et à 0 les autres bits. Les signatures virtuelles sont construites pour chaque mot-clé, pour chaque document de la collection où ce mot-clé se produit. Or, la plupart des bits de la signature sont souvent à 0, vu qu'un terme n'apparaît pas dans tous les nœuds d'un document en même temps, mais souvent, dans quelques uns uniquement. Pour éviter d'enregistrer toute la signature dans le fichier inverse étendu, on ne sauvegarde que les positions des bits qui sont à 1 dans la signature virtuelle. En recherche d'information semi-structurée, le poids d'un terme exprime son importance dans le nœud qui le contient. Ce poids est calculé grâce à de nombreuses formules telles que *ief* (*Inverse Element Frequency*) [5], ou *tf-idf* (*Term Frequency - Inverse Tag and Document Frequency*) [10]. On retrouve dans [8] une évaluation de plusieurs formules de pondération, et les expérimentations ont démontré que la formule la plus optimale est *tf-ief* (*term frequency - inverse element frequency*) et c'est celle-ci que nous utiliserons. Donc, le poids  $w_{i,j}$  du terme  $t_i$  dans la feuille  $f_j$  est calculé avec la formule normalisée de *tf-ief* suivante :

$$w_{i,j} = tf_{i,j} * ief_i = \frac{nbocc(t_i, f_j)}{nbterms(f_j)} * \log\left(\frac{|F_c|}{|nf_j|}\right) / \log(|F_c|)$$

Où :  $|F_c|$  est le nombre total de nœuds feuilles dans la collection et  $|nf_j|$  le nombre de nœuds feuilles contenant le terme  $t_i$ .

#### 4.4. Procédure de recherche dans notre index

Il existe deux types de requêtes dans la recherche d'information semi-structurée : les requêtes orientées contenu (Content Only) ou orientées contenu et structure (Content And Structure). L'index que nous proposons permet de répondre à tous les types de requêtes CO ou CAS.

##### 4.4.1. Requête de type CO

```

Program SearchCo(Liste of node)
begin
  Résultat := ensemble vide ;
  chercher le mot-clé demandé dans le fichier inverse et récupérer
  les couples (docID, nodeliste)
  For chaque couple retourné Do
  begin
    For chaque couple (nodepos, w) de nodeliste Do

```

```

begin
  extraire du DataGuide annoté le chemin d'étiquettes c menant
  au noeud Ndga annoté contenant l'intervalle comportant nodepos
  et la valeur minpos de cet intervalle ;
  posndb=nodepos-minpos+1; {la position du noeud Ndb de la base
  de données accessible par le chemin c}
  insérer la ligne (docID,c,w,posndb) dans le résultat ;
End;
End;
  trier le résultat par ordre décroissant des w ;
end ;

```

#### 4.4.2. Requête de type CAS

```

Program SearchCasI(Liste of node)
begin
  résultat := ensemble vide ;
  extraire du DataGuide annoté tous les chemins correspondant au
  chemin de la requête ;
  for chaque chemin c retourné do
  begin
    récupérer l'intervalle contenu dans le noeud du DataGuide
    accessible c ;
    chercher dans le fichier inverse le mot-clé t de la requête et
    récupérer la liste des couples de type (docID, nodeliste)
    correspondants ;
    for chaque couple retourné do
    begin
      for chaque couple de type (nodepos,w) de nodeliste tel que
      nodepos n'appartient pas à l'intervalle récupéré do
      begin
        posndb :=nodepos-minpos+1 ; {la position du noeud Ndb de la
        base de données accessible par le chemin c avec}
        insérer la ligne (docID,c,w,posndb) dans le résultat. ;
      end;
    end;
    trier le résultat par ordre décroissant des w ;
  end;
end;

```

## 5. Tests et résultats

Nos expérimentations ont été menées sur cinq collections de documents semi-structurés. Les index résultants permettent de répondre efficacement aux requêtes de contenu et de structure. Ci-dessous sont exprimées les tailles des index :

**Table 1 : Résultats de l'expérimentation**

	Reuters	Religion	Shakespeare	DBLP	NASA
Nombre de documents	23	4	37	32	1
Taille de la collection (1)	27,3	7	7,54	130	23,8
Taille de l'index (2)	17,21	5	6,77	92,06	9,07
(2) / (1)	0,63	0,71	0,89	0,70	0,38

## 6. Conclusion

Nous avons présenté dans cet article une approche d'indexation des documents semi-structurés basée sur le *DataGuide*. L'approche réalise un lien virtuel entre le *DataGuide* et l'index de contenu qu'est le fichier inverse étendu, elle permet ainsi l'indexation du contenu et de la structure. Sa force principale est la façon dont le lien entre ses deux structures est établi en utilisant la signature virtuelle qui n'est calculée qu'au moment de l'indexation et ne fait pas partie de l'index lui-même. Notre index permet de répondre aux requêtes de contenu et de structure et ceci quelque soit le type de structure de la requête. Il est donc flexible du moment qu'il permet l'utilisation de n'importe quelle méthode de pondération et peut indexer des collections hétérogènes. Nous avons réalisé un système de recherche d'information basé sur cet index, ce dernier a été testé et permet de retourner de bons résultats. La prochaine étape de notre travail consistera à évaluer notre index sur un corpus issu de la campagne d'évaluation INEX.

## 7. Bibliographie

1. Chen Y and Aberer K. Combining Pat-Trees and Signature Files for Query Evaluation in Document Databases. In Database & Expert Systems, 1999.
2. Cooper B and Shadmon M. The Index Fabric: Technical Overview. Technical report, RightOrder Inc., 2001.
3. Dongwook S, Hyuncheol J and Honglan J. BUS: An Effective Indexing and Retrieval Scheme in Structured Documents. Proceedings of the 3rd ACM, 1998.
4. Goldman R and Widom J. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. Proceedings of the 23rd VLDB, 1997.
5. GRABS T., SCHECK H.-J., Flexible information retrieval from XML with PowerDB XML, Proceedings in the First INEX Workshop, 2002, p. 26-32.
6. McHugh J., Abiteboul S., Goldman R., Quass D. and Widom J. Lore: A DB Management System for Semistructured Data. SIGMOD Rec., 26(3):54-66, 1997.
7. Qinghua Z, Shaorong L, Wesley W.C, Ctree: A Compact Tree for Indexing XML Data, Computer Science Department Univ of California. ACM, WIDM'04, 2004.
8. Karen Sauvagnat, Mohand Boughanem, Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée, CORIA 2006.
9. F. Weigel, H. Meuss, F. Bry, and K. U. Schulz, Content-Aware DataGuides: Interleaving IR and DB Indexing Techniques for Efficient Retrieval of Textual XML Data. In Proc. 26th Europ. Conf. on IR, 2004.
10. ZARGAYOUNA H., Contexte et sémantique pour une indexation de documents semi-structurés, Actes de CORIA 04, Toulouse, France, 2004, p. 161-178.
11. J. Zobel, A. Moffat, and K. Ramamohanarao. Inverted Files Vs Signature Files for Text Indexing. ACM Transactions on DB Systems, 23(4):453-490, 1998.



## La réingénierie de l'ontologie : vers une nouvelle approche

Ghizlène Soulimane, Mimoun Malki \*, Sidi Mohamed Benslimane \*

\* Laboratoire EEDIS Département Informatique, UDL, Sidi Bel Abbes. ALGERIE  
[ghizlenes@yahoo.fr](mailto:ghizlenes@yahoo.fr) , [malki@univ-sba.dz](mailto:malki@univ-sba.dz) , [benslimane@univ-sba.dz](mailto:benslimane@univ-sba.dz)

**Résumé :** Les ontologies visent à capturer la connaissance du domaine et fournissent un consensus sur la compréhension commune du domaine qui peut être réutilisée et partagée à travers des applications, des groupes d'ingénieurs ou d'agents logiciels. Après une certaine période ou selon des besoins applicatifs, l'ontologie change, elle évolue et doit être modifiée en appliquant des changements conceptuels et/ou représentationnels pour refléter au mieux le monde réel qu'elle représente.

Nous nous sommes intéressés à l'évolution de l'ontologie sous un nouveau jour et ceci en appliquant des techniques de réingénierie. Le principe soutenu dans l'approche proposée dans cet article est que nous pouvons construire une nouvelle ontologie, dans un langage de représentation des plus récents tel que OWL, en utilisant les connaissances capturées par une ancienne ontologie représentée dans un ancien langage tel que DAML+OIL et ceci par un processus de réingénierie qui conduit à une représentation intermédiaire, dans notre cas, c'est UML qui est utilisé comme modèle conceptuel qui sera mis à jour par l'ontologiste avec la présence d'un expert du domaine et d'un utilisateur.

**MOTS-CLÉS :** ontologie, construction d'ontologies, réingénierie, rétro-ingénierie, évolution, DAML+OIL, UML, OWL.

### 1 Introduction

Les ontologies visent à capturer la connaissance d'un certain domaine et fournissent un consensus sur la compréhension commune de ce domaine qui peut être réutilisée et partagée à travers des applications ou des groupes. Après une certaine période ou selon des besoins applicatifs, l'ontologie doit être modifiée pour refléter au mieux le monde réel. Cette modification se fait par l'application de plusieurs changements sur l'ontologie.

Un nouveau domaine de recherche a vu le jour il y a une vingtaine d'années dans le milieu académique, en l'occurrence la réingénierie, dans le but justement de reconsidérer les ontologies existantes dans une nouvelle perspective, en migrant vers de nouvelles technologies.

Le domaine de la réingénierie des ontologies est un domaine nouveau, ce qui explique le nombre limité des approches s'attaquant à cette nouvelle perspective. Nous aurons l'occasion de voir deux approches dans cet article : celle de la réingénierie des ontologies (élaborée par le groupe d'Ontologie du laboratoire d'intelligence artificielle à UPM [2], [3] et plus récemment, l'approche de El-Ghalayini et al [4] étudie la possibilité de générer un modèle conceptuel de domaine par un processus de rétro-ingénierie d'ontologie en utilisant un modèle pivot le BWV. Mais la question qui nous est parue évidente est : « Pourquoi utiliser un modèle pivot ? Est-ce que le passage direct (ontologie/schéma conceptuel) n'est pas possible ? Est-ce que la sémantique englobée dans l'ontologie n'est pas altérée ? »

Ce papier est organisé comme suit : dans la section 2, nous discutons quelques travaux dans le domaine de la réingénierie des ontologies. La section 3 expose l'architecture proposée dans notre approche et la section 4 détaille notre approche de réingénierie des ontologies. Enfin, la section 5 contient une discussion et des perspectives.

### 2 Travaux antérieurs

La réingénierie comporte deux notions, d'une part celle de l'ingénierie directe (construction d'ontologie) où apparaissent plusieurs approches, méthodes et méthodologies ; ces dernières ont été classées selon plusieurs critères [5].

D'autre part, la notion de rétro ingénierie des ontologies où nous avons recensé une seule approche, celle de El-Ghalayini et al [4] où la rétro ingénierie d'une ontologie DAML+OIL vers un modèle

conceptuel se fait en passant par un modèle pivot appelé le modèle BWW ('Bunge-Wand-Weber model')[6], appelé Mapping des constructions d'ontologie aux constructions de modélisation conceptuelle.

L'approche de Gomez-Perez [2] emploie l'ensemble de représentations intermédiaires proposées par METHONTOLOGY pour aboutir au modèle conceptuel. En fait, cette méthodologie de réingénierie peut être considérée comme une extension du cadre de METHONTOLOGY. Frederik Gailly et Geert Poels ont eux aussi utilisé METHONTOLOGY comme base pour développer leur approche de réingénierie d'ontologie de domaine d'affaires [7].

Et dans la même lancée S. H. Kang et S. Lau [8] proposent une approche de réingénierie des ontologies basée sur celle de Gomez-Perez [3]. Ils ajoutent un processus de révision à l'étape de restructuration dans le schéma global de l'approche de réingénierie.

### **3 Approche proposée pour la réingénierie des ontologies :**

#### **3.1 Motivations**

Un des objectifs de la réingénierie ontologique est l'évolution de l'ontologie en question. Il y a plusieurs cas où l'évolution d'ontologie est applicable. Dans la classification de la terminologie concernant le domaine de la réingénierie de E. J. Chikofsky et J. H. Cross [9], " la réingénierie correspond au processus d'examen et de modification d'un système dans le but de le reconstituer dans une nouvelle forme puis l'implémenter. ", et ceci correspond au but ultime de notre approche qui est de faire la réingénierie d'une ontologie de domaine issue d'un système légué (écrite en langage DAML), vers une nouvelle ontologie de domaine qui englobe les changements survenus (modification) dans ce même domaine et qui sera écrite dans un langage plus récent (le langage OWL). Un changement de représentation (DAML/ OWL dans notre cas) n'est pas toujours un changement de spécification. Si une ontologie est traduite dans un langage différent avec une expressivité équivalente ou plus grande, la spécification ne change pas. Si l'expressivité du langage cible est plus limitée, donc le changement de représentation implique un changement de spécifications. [1].

#### **3.2 Architecture proposée**

L'architecture proposée est composée de trois modules , le premier est responsable de l'extraction des entités ontologiques de l'ontologie léguée représentée dans notre étude en DAML+OIL, le deuxième module est un premier moteur de transformation responsable du mapping des entités ontologiques (résultat de la première étape) en constructeurs conceptuels UML, ce moteur de transformation contient le premier ensembles de règles de mapping ; enfin, le troisième module de cette architecture est un deuxième moteur de transformation qui fait le mapping des constructeurs conceptuels UML en entités ontologiques (OWL).

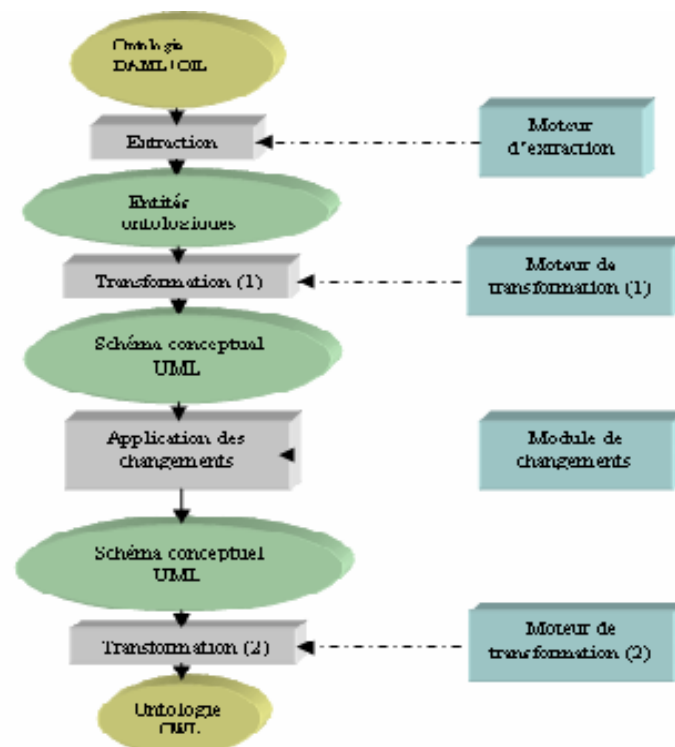


Fig1 : Processus de ré-ingénierie d'une ontologie DAML+OIL vers OWL

## 4 Processus de réingénierie des ontologies (DAML-UMLOWL)

### 4.1-Extraction des entités ontologiques de l'ontologie léguée (DAML+OIL) :

Le but de cette phase est de faire ressortir de l'ontologie léguée ses composants ontologiques concepts, relations entre concepts, propriétés et contraintes (toutes les classes, propriétés (datatype et object property), restrictions).

### 4.2-Mapping des entités ontologiques en constructeurs conceptuelsUML :

Pour faire ressortir les similarités entre DAML et UML, Kenneth Bacławski et ses collègues [10] ont présenté un mapping de haut niveau (High-Level mapping) présenté sur la figure suivante.

Concept DAML	Concepts UML similaires
Ontologie	Package
Classe	Classe
Ensemble (union, disjoint)	Difficiles à représenter
Hierarchie	Relation de généralisation entre classes
Propriété	Attributs, associations et classes
Hierarchie	Pas pour les attributs, généralisation limitée aux associations et aux classes
Restriction	Contraint les fins d'association (cardinalité), et implicitement les attributs de classes.
Data type	Type de données
Instances et valeurs	Instances d'objet et valeurs d'attributs.

Fig 2: Mapping de haut niveau DAML-UML de Backlawski

Le mapping utilisé dans notre approche est inspiré de l'outil DUET [11] DAML UML Enhanced Tool (DUET) du projet CODIP qui fournit un environnement basé UML pour la visualisation, développement

et manipulation des ontologies DAML, mais a été redéfini pour représenter une ontologie DAML+OIL sous la forme la plus connue et utilisée en UML : le diagramme de classe.

### Règles de transformation DAML+OIL /UML

Le premier moteur de transformation contient cet ensemble de règles :

R1 : tout concept dans l'ontologie est représenté par une DAMLClass et sera représenté par une classe UML de même nom.

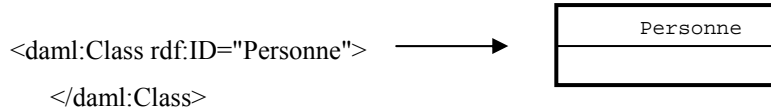


Fig 3 : Classe UML Personne.

R2 : Une Restriction sur une objectProperty devient une association ou un rôle UML. La classe source de l'association est une subclass de la restriction, commençant par onProperty pour l'objectProperty avec le toClass qui devient la classe cible de l'association. Le nom de l'Object Property devient le nom de rôle du toClass classe.

Par exemple les clauses suivantes :

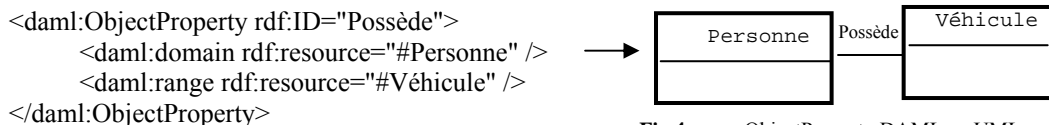


Fig 4 : une ObjectProperty DAML en UML.

<daml:Class rdf:ID='vehicule'>  
</daml:Class>

R3 : une ObjectProperty avec un domaine et un range devient une association UML. L'association a la classe domaine comme classe source pour l'association UML et le range comme classe cible.

R4 : Une DatatypeProperty avec un domaine et un range est représenté en un attribut d'une classe UML. Le nom de l'attribut est celui de la propriété, la classe où il figure est le domaine de cette propriété, et son type est le range de la propriété.

R5 : les sous-classes de l'ontologie deviennent une généralisation UML.

R6 : les cardinalités dans des restrictions dans une ontologie daml se traduisent par des multiplicités dans les associations UML.

### 4.3-Eventuellement MAJ des données conceptuelles pour représenter le domaine.

Cette partie de l'approche est optionnelle et elle implique la présence des utilisateurs de l'ontologie et d'un expert du domaine pour se mettre d'accord sur les changements à apporter à la nouvelle conceptualisation de l'ontologie, cette mise à jour peut être par ajout/suppression ou modification de concepts ou propriétés (relation) présents dans la dernière conceptualisation représentant la première ontologie et ceci dans le but de la faire évoluer non seulement dans sa représentation (langages d'un ancien DAML+OIL à un plus récent OWL) mais aussi dans sa conceptualisation.

### 4.4-Transformations de ces constructeurs en entités ontologiques (OWL).

La perspective que UML devienne un langage de modélisation conceptuelle pour Web Sémantique est devenu un sujet de recherche très en vogue.

Les avantages de UML comme un langage de modélisation d'ontologie ont été soulevés par Baklawski et al dans [10].

#### **Règle de mapping UML en OWL :**

Voici l'ensemble de règles de transformation constituant le deuxième moteur de transformation :

R1 : OWL et UML sont basés sur les classes, alors une classe UML est tracée en une classe OWL en lui assignant un nom au type approprié.

Exemple: `<owl: class rdf: ID="Personne"/>` est la transformation de la classe représentée en figure 2

R2 : il y a deux sources de propriétés dans UML : En premier lieu, une classe UML qui possède ses propres (attributs) ; si un attribut propriété d'une classe UML est une autre classe, cet attribut se transforme en une owl :ObjectProperty, sinon il est transformé en une owl :DataTypeProperty.

R3 : une instance d'association binaire UML est traduite directement en une owl :ObjectProperty. (Le domaine et le rang sont les deux classes de l'association).

R4 : Dans UML, une classe peut exister comme généralisation d'une ou plusieurs autres classes. L'élément de généralisation est le synonyme du constructeur owl :SubClassOf. La traduction UML-OWL est directe.

La classe Homme est sous-classe de la classe personne.

```
<owl: class rdf: about="Homme">  
<owl: subClassOf rdf: resource="#Personne" />  
</owl: class>
```

R5 : Dans OWL une propriété appliquée à une classe peut être contrainte par des restrictions de cardinalité donnant le nombre d'instances pouvant participer à la relation. En UML, une association peut avoir une cardinalité min et max spécifiée aux extrémités. OWL donne la possibilité de déclarer des paires de valeurs individuelles (objectproperty) une l'inverse de l'autre, ainsi, si une association binaire UML a une multiplicité à une extrémité, la propriété correspondante OWL aura la même multiplicité. [12]

## **5 Discussions et perspectives**

Nous avons vu plus haut qu'il existe une approche de réingénierie des ontologies, celle du groupe d'ontologie du laboratoire d'intelligence artificielle à UMP [2]. C'est une extension de la méthode METHONTOLOGY. Cette dernière inclut un cycle de vie basé sur les prototypes d'évolution, et laisse ajouter, changer, enlever des termes de chaque nouvelle version (prototype) de l'ontologie en question jusqu'à arriver à une ontologie qui comble les attentes des utilisateurs. Ce sont ces versions qui sont utilisées dans le processus de re-ingénierie des ontologies ; Donc, cette méthodologie de réingénierie n'est applicable qu'aux ontologies construites suivant le cycle de vie présenté par METHONTOLOGY et seulement par cette méthodologie.

El-Ghalayini et al. [4] proposent dans leur approche de rétro-ingénierie d'utiliser le modèle BWW comme modèle pivot pour la phase de rétro-ingénierie qui correspond au passage DAML+OIL vers UML. C'est justement l'utilisation de ce modèle pivot qui a engendré notre problématique : est-ce que le passage direct DAML+OIL vers UML n'est pas possible ? L'idée développée dans notre travail est que UML peut jouer le rôle du modèle pivot au même titre que le modèle BWW, remplaçons, donc, le modèle BWW par le modèle UML ce qui nous donne le passage direct de DAML+OIL vers UML (UML constituant en même temps le modèle pivot et le modèle conceptuel utilisé pour décrire l'ontologie)

Voyons maintenant plus en détail les règles de transformation utilisées dans chacune des deux approches (El-ghalayini et al rétro-ingénierie de notre approche).

Approche de El-Ghalayini et al. 2005

Notre approche 2007

DAML	BWW	UML		DAML	UML
Classe	Classe, sorte, sorte naturelle	Type d'entité		Classe	Classe
Data Type property	Propriété intrinsèque	Attribut d'entité		Datatype property Domaine + porté	Attribut de classe
Propriété « Super/ sous-classe »	Propriété mutuelle « Naturelle/ sous-sortes »	relation « Généralisation/ Spécialisation »		Object property Domaine+range « Sous –propriété de « Sous-classe de »	Association binaire « Généralisation/ Spécialisation »
Axiome et restriction	Loi	Contrainte sémantique (Type de relation, nombre et type d'entité)		* Restriction sur une objet property * Restriction sur une classe	* Association ou Rôle d'association * Cardinalité d'association

**Fig 5** : tableau comparatif des règles de transformation DAML/UML des deux approches (El-Ghalayini et al.2005 et la notre).

Notre approche parvient aux mêmes résultats de transformation en utilisant UML comme modèle pivot (remplaçant le modèle BWW) et comme modèle conceptuel représentant la conceptualisation de l'ontologie héritée écrite en DAML.

Notons que l'approche de El-Ghalayini et al ne fait que la partie rétro-ingénierie d'une ontologie DAML, alors que celle présentée dans cet article termine tout le processus de réingénierie ontologique en construisant une ontologie OWL à partir des données conceptuelles UML.

C'est avec l'implémentation de l'idée d'une approche de réingénierie d'ontologie que plusieurs perspectives sont apparues intéressantes :

Le système implémenté actuellement ne permet de prendre en charge que les changements que nous avons appelés représentationnels (DAML+OIL vers OWL), et sera plus efficace lors de l'implémentation des changements conceptuels identifiés dans la partie approche, en incorporant un lexicon tel que WordNet et un calcul de distance sémantique (ou plusieurs) à notre architecture pour enrichir le modèle conceptuel avec de nouveaux concepts et relations, ou encore pour garder le modèle conceptuel cohérent lors de suppression ou de déplacement de concepts ou de relations.

Travailler avec les instances de l'ontologie héritée est notre prochain objectif pour englober tous les composants d'une ontologie même si ces instances n'apparaîtront pas dans le schéma conceptuel UML puisque ce dernier est représenté par un diagramme de classe mais seront utilisés lors de la deuxième phase, la construction de la nouvelle ontologie.

## Bibliographie

1. M. Klein. Change Management for Distributed Ontologies. PhD thesis, Department of Computer Science, Vrije Universiteit Amsterdam, 2004.
2. Gomez-Perez, A. (1999) Applications of Ontologies and Problem-Solving Methods, in Proceedings of Workshop at Thirteenth Biennial European Conference on Artificial Intelligence, 1 - 4 Oct., Berlin, Germany.
3. Gomez-Perez, A., Fernandez-Lopez, M. and Corcho, O. (2004) Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and Semantic Web, Springer, London, Berlin, Heidelberg, New York, Hong Kong, Milan, Paris, Tokyo
4. Haya El-Ghalayini, Mohammed Odeh, Richard McClatchey & Tony Solomonides ;' REVERSE ENGINEERING ONTOLOGY TO CONCEPTUAL DATA MODELS ' Centre for Complex Cooperative Systems, CEMS Faculty, University of the West of England (UWE), Coldharbour Lane, Frenchay, Bristol BS16 1QY, United Kingdom.2005.
5. Ghizlène Soulimane, Djelloul Bouchiha, Sidi Mohamed Benslimane : La réingénierie des ontologies: Etat de l'art Conférence Internationale sur l'Informatique et ses Applications CIIA06 15-16 Mai 2006

6. Peter Fettke, Peter Loos, 'ONTOLOGICAL EVALUATION OF REFERENCE MODELS USING THE BUNGE-WAND-WEBER MODEL' Lehrstuhl für Wirtschaftsinformatik und BWL Johannes Gutenberg-University Mainz 2003.
7. Frederik Gailly Geert Poels Towards Ontology-driven Information Systems: Redesign and Formalization of the REA-ontology, 25 April 2007 <http://users.ugent.be/~fgailly/phd/docs/BIS2007-fgailly.pdf>
8. Kang, SH and Lau, SK, Ontology Revision on the Semantic Web: Integration of belief revision theory, 40th Annual Hawaii International Conference on System Sciences (HICSS 2007), Waikoloa, Hawaii, USA, January 2007, 61. Copyright 2007 IEEE.
9. Chikofsky E.J., Cross II J.H.: *Reverse engineering and design recovery: A taxonomy*. *IEEE Software*, 13, 1990.
10. Kenneth Baclawski, Mieczyslaw M. Kokar, Paul A. Kogut, Lewis Hart, Jeffrey E. Smith, William S. Holmes III, Jerzy Letkowski, and Michael L. Aronson. Extending UML to support ontology engineering for the semantic web. In *UML Concepts, and Tools, 4th International Conference*, volume 2185 of *LNCS*, pages 342–360, Toronto, Canada, October 2001
11. [http://codip.grci.com/wwwlibrary/wwwlibrary/DUET\\_Docs/DUET\\_UsersGuide\\_V0.3.htm](http://codip.grci.com/wwwlibrary/wwwlibrary/DUET_Docs/DUET_UsersGuide_V0.3.htm)
12. Benslimane, S. M., Benslimane, D., Malki, M., Maamar, Z., Thiran, P., Amghar, Y. and Hacid, M. : "Ontology Development for The Semantic Web : An HTML Form-based Reverse Engineering Approach". *Journal of Web Engineering (JWE)*, Vol. 6, No.2, 97-199, June, 2007.

## Appariement ontologique

A.Boubekeur<sup>1</sup>, M. Malki<sup>2</sup>

<sup>1</sup>Département d'Informatique Université de Tiaret, Algérie

[Boub68a@yahoo.fr](mailto:Boub68a@yahoo.fr)

<sup>2</sup>Département d'Informatique Université de Sidi Bel-Abbes, Algérie

[Malki\\_m@yahoo.com](mailto:Malki_m@yahoo.com)

**Résumé :** La grande diversité des ontologies sur un même domaine est en soi une source d'hétérogénéité. Une solution à cette difficulté se situe dans le développement de techniques de coordination et d'appariement d'ontologies. Dans ce cadre, l'appariement des ontologies, permettant d'établir des liens sémantiques (équivalence, spécialisation,...) entre ontologies du même domaine, est une autre problématique centrale afin de pouvoir manipuler de façon conjointe plusieurs ontologies au sein d'une même application, ou de permettre à deux applications utilisant des ontologies différentes de communiquer. Notre focalisation est sur la composition d'une variété de mesures de similarité entre les entités des ontologies OWL et sur un processus de calcul itératif pour ressortir les K-Meilleurs termes (concepts, relations, propriétés.....) appariés d'une multiplicité unidirectionnelle de type 1 à n.

**Mots clefs:** Ontologies, interopérabilité sémantique, alignement, mesure de similarité, Word Net, OWL, XML

### 1 Introduction

L'appariement d'ontologies permet les échanges de manière sémantique. Cet appariement peut se situer au niveau sémantique et conceptuel ou au niveau des instances et des données [13]. Les travaux existant pour établir les relations de mise en correspondance ou d'appariement se focalisent à la fois sur les aspects d'automatisation et de découverte de ces relations et aussi sur l'aspect formalisation pour le raisonnement. Ces approches se basant sur des heuristiques permettent d'identifier des similarités syntaxiques et structurelles. L'absence de structure commune, de similarités syntaxiques et d'instances communes rend le problème de mise en correspondance difficile et souvent basé sur de fortes hypothèses pas toujours très réalistes. Le problème principal auquel est confronté le processus de d'appariement est lié à l'hétérogénéité sémantique introduite quand différents modèles sont construits avec différentes terminologies pour représenter la même information dans les différentes sources.

Les ontologies sont la solution pour parvenir à l'interopérabilité sémantique sur le Web. Et pourtant, de plus en plus, on se rend compte qu'on n'arrivera pas toujours à un consensus. Au contraire, un grand nombre d'ontologies verra le jour sur un même sujet, mais elles diffèrent, non seulement au niveau des choix terminologiques ou de la langue choisie, mais également au niveau sémantique et conceptuel : il sera nécessaire d'apparier les termes (concepts, relations, propriétés, ...) sous-jacents; parce que les points de vue seront différents.

La découverte de similarités repose sur [12] : Les techniques terminologiques permettent d'exploiter toute la richesse des noms des termes, en particulier dans les domaines où les homonymes sont rares. Les techniques structurelles permettent de trouver des mises en correspondance potentielles du fait de la position dans la taxonomie des termes candidats lorsque l'exploitation de leur syntaxe ne suffit pas. Enfin, lorsque terminologiquement, ou structurellement, il n'est pas possible de trouver de correspondances, les techniques sémantiques montrent tout leur intérêt [12].

Notre travail se veut d'aborder ces techniques et de fournir un procédé multi-appariement d'ontologies OWL. Le but est, et en utilisant WordNet [6], de trouver pour chaque terme (concept, relation, attribut,...) de l'ontologie source de quels termes de l'ontologie cible il peut être similaire ou sémantiquement rapproché (ceux avec lesquels ils partagent des généralisants) puis d'identifier le terme dont il est le plus proche. Notre démarche consiste donc à définir un seuil et procède par la suite à des calculs de similarité entre les différents termes. A partir de ce seuil, on déterminera les entités à considérer comme similaires. S'il s'agit de distance entre entités, les valeurs inférieures au seuil indiqueront une quasi-équivalence des entités, tandis que pour les mesures de similarité, les valeurs supérieures au seuil indiquent que les entités sont similaires.



La section 2 de ce papier présente un état de l'art sur les travaux relatifs. La sections 3 décrit notre processus d'appariement des ontologies. En section 4, nous montrons les résultats expérimentaux et les évaluations de notre stratégie proposée. La dernière section fait des suggestions de conclusion pour les travaux futurs.

## 2 Etat de l'art

Il y a beaucoup d'algorithmes différents qui implémentent l'appariement. Ces algorithmes peuvent être généralement classés le long de deux dimensions. D'une part il y a la distinction entre l'assortiment qui se base sur le schéma de l'ontologie et celui qui se base sur les instances de l'ontologie [13]. Un assortiment basé sur le schéma prend les différents aspects de concepts et de relations des ontologies et emploie une certaine mesure de similarité pour déterminer la correspondance [9]. Un assortiment basé sur les instances prend les instances qui appartiennent aux concepts dans les différentes ontologies et compare ces dernières pour découvrir la similarité entre les concepts [10]. D'autre part, il y a la distinction entre l'assortiment au niveau élément et l'assortiment au niveau structure. Un appariement au niveau élément compare les propriétés particulières d'un concept ou d'une relation, telle le nom, et emploie ces dernières pour trouver des similarités [7]. Un appariement niveau structure compare les structures (par exemple, la hiérarchie de concept) des ontologies pour trouver des similarités [7], [13]. Pour une classification plus détaillée des techniques d'alignement nous nous référons [13]. Ces techniques peuvent également être combinées [13]. Par exemple, Ancre-Prompt [7], rapprochent des ontologies vues comme des graphes au sein desquels les noeuds sont des classes et les liens sont des propriétés. Le système prend en entrée un ensemble d'ancres qui sont des couples d'éléments liés. Un algorithme analyse les chemins dans le sous graphe délimité par les ancres et détermine quelles classes apparaissent fréquemment dans des positions similaires sur des chemins similaires. Ces classes correspondent ainsi vraisemblablement à des concepts similaires [7]. Maedche et Staab, dans [5], proposent d'exploiter les relations entre les concepts. Des travaux de Ehrig et Staab sont focalisés sur l'efficacité des algorithmes de génération de mappings [3]. Des heuristiques pour restreindre le nombre de mappings candidats sont utilisées et une ontologie permet de classer les mappings candidats selon qu'ils sont plus ou moins prometteurs. Dans le travail de Maedche et Staab, une mesure globale de similarité entre deux hiérarchies est calculée, consistant à comparer les éléments parents et fils de tous les éléments communs [5].

Cette méthode n'est adaptée que si les hiérarchies sont bien structurées et contiennent de nombreux éléments en commun. Enfin Euzenat et Valtchev dans [4] proposent une mesure de similarité dédiée aux ontologies décrites en OWL-Lite, permettant d'agréger différentes techniques de comparaison exploitant les constructeurs de OWL-Lite dans une mesure commune.

D'autres travaux exploitent les données associées aux ontologies et appliquent des techniques d'apprentissage automatique sur ces données [2]. Enfin, des travaux réalisés à l'université de Trento en Italie mettent l'accent sur les aspects sémantiques et proposent d'utiliser WordNet pour aider à mettre en correspondance sémantiquement des éléments d'une hiérarchie de classification [1]. Notre approche se différencie des travaux susmentionnés par le mécanisme itératif de la mise en correspondance entre termes (concepts et relations) présents dans les ontologies d'entrée et par la composition de multiples mesures de similarité. Cette composition aide à raffiner les résultats obtenus au préalable. Notre étude traite d'une part, l'hétérogénéité syntaxique en définissant des moyens pour représenter ces ontologies dans un seul langage, langage pivot, le standard OWL, et d'autre part, l'hétérogénéité sémantique qui se base sur l'interrogation d'une ressource linguistique, telle le WordNet, tout en l'accompagnant d'un degré de similarité appelé le degré de tolérance qui indique la plausibilité de cet appariement. Notre étude définit donc techniquement le processus d'appariement de plusieurs ontologies et de son résultat : multi-alignement.

## 3 Expérimentations

Nos études de cas ont été réalisées avec l'intention d'évaluer leurs comportements devant un traitement d'ontologies avec un grand nombre de termes et de structures taxonomiques bien différentes. Le choix des ontologies utilisées a été donné à la disposition des ontologies d'institutions et de contexte connu dans des

domaines d'applications spécifiques. Ces ontologies sont de la compagnie I3CON<sup>2</sup>. I3CON, le premier effort dans la communauté de la recherche de l'intégration et de l'interprétation de l'ontologie et du schéma, est pour fournir des outils et un framework d'évaluation systématique de l'intégration des ontologies et des schémas. I3CON fournit 10 paires d'ontologies accompagnées de leur fichier de référence qui contient les correspondances déterminées manuellement entre deux ontologies.

### 3.1 Des mesures pour l'évaluation des résultats

Pour évaluer la précision, l'efficacité et la performance de notre algorithme d'appariement, nous employons les mesures de *précision*, de *rappel* et de *f-mesure*.

Le *rappel* est la proportion de correspondances correctes renvoyées par l'algorithme parmi toutes celles qui sont correctes (en incluant aussi des correspondances correctes que l'algorithme n'a pas détectées). Le *rappel* mesure l'efficacité d'un algorithme. Plus la valeur de *rappel* est élevée, plus le résultat de l'algorithme couvre toutes les correspondances correctes.

La *précision* est la proportion des correspondances correctes parmi l'ensemble de celles renvoyées par l'algorithme. Cette mesure reflète la précision d'un algorithme. Plus la valeur de *précision* est élevée, plus le bruit dans le résultat de l'algorithme est réduit, et donc plus la qualité de résultat est imposante.

Enfin, la *f-mesure* est un compromis entre le *rappel* et la *précision*. Elle permet de comparer les performances des algorithmes par une seule mesure. La *f-mesure* est définie par

$$f\text{-mesure} = (2 * \text{rappel} * \text{précision}) / (\text{rappel} + \text{précision}) \quad (1)$$

Nous utilisons aussi la mesure globale (overall measure) définie dans [14]. Cette mesure, appelée l'exactitude, correspond à l'effort exigé pour corriger le résultat renvoyé par l'algorithme afin d'obtenir le résultat correct. La mesure globale est toujours inférieure à la *f-mesure*. Elle n'a un sens que dans le cas où la *précision* n'est pas inférieure à 0,5, c'est-à-dire si au moins la moitié des correspondances renvoyées par l'algorithme sont correctes. En effet, si plus de la moitié des correspondances sont erronées, il faudrait à l'utilisateur plus d'effort pour enlever les correspondances incorrectes et d'ajouter les correspondances correctes mais absentes, que pour mettre en correspondance manuellement les deux ontologies à partir de zéro.

$$\text{Overall} = \text{rappel} * (2 - 1/\text{précision}) \quad (2)$$

Nous notons :

- $F$  – le nombre des correspondances renvoyées par l'algorithme
- $T$  – le nombre des correspondances correctes déterminées manuellement par des experts (celles dans le fichier de référence)
- $C$  – le nombre des correspondances correctes trouvées (appelé aussi vrais positifs)
- Ainsi,  $\text{précision} = C / F$  ;  $\text{rappel} = C / T$  ;

### 3.2 Méthodologie d'évaluation

Pour chaque paire d'ontologie, notre outil d'appariement appelé OntAlign est exécuté avec des seuils différents de similarité. Le seuil est utilisé dans la stratégie pour déterminer des correspondances considérées comme correctes. La valeur de similarité totale de deux entités appartenant aux ontologies, calculée par des techniques terminologique et sémantique, est comparée avec le seuil. Deux entités (concepts ou relations) sont considérées comme similaires si leur valeur de similarité dépasse ce seuil. Le résultat final de l'algorithme (la liste des correspondances entre deux ontologies) dépend donc du seuil choisi. Plus le seuil de similarité est élevé, plus la qualité du résultat est élevée (la *précision* augmente, le nombre des correspondances incorrectes trouvées diminue) mais plus des correspondances correctes sont considérées comme incorrectes. En continuant à augmenter le seuil de similarité, les valeurs de *f-mesure* et d'*overall* diminuent car le nombre des correspondances correctes mais pas trouvées devient très important. La raison est qu'il y a des entités qui représentent des concepts similaires mais leurs définitions

<sup>2</sup> Information Interpretation and Integration Conference; <http://www.atl.lmco.com/projects/ontology/i3con.html>.

dans des ontologies sont un peu différentes, donc leur valeur de similarité, calculée en se basant sur leurs définitions, n'atteint pas une valeur élevée (le maximum est 1.0), et donc ne dépasse pas le seuil.

Les correspondances trouvées par notre algorithme sont comparées avec les correspondances correctes dans le fichier de référence (fourni aussi par la campagne de test I3CON), pour calculer le nombre de correspondances correctes et incorrectes trouvées par l'algorithme. Les valeurs des mesures d'évaluation (précision, rappel, f-mesure, overall) sont calculées pour chaque cas et affichées dans un tableau.

#### ♦ Evaluation de paires d'ontologies dans la campagne I3CON

Les Tableaux tabl.1 et tabl. 3 résument des informations concernant les paires d'ontologies de test, avec le nombre de concepts, de relations, de propriétés, d'instances dans les deux ontologies, la ressource externe interrogée et le nombre de correspondances correctes T dans le fichier de référence fourni.

##### ♦ La paire « Animals » ( $O_A:animalsA.owl$ - $O_B:animalsB.owl$ )

**Tabl. 1.** Informations concernant la paire d'ontologies « Animals »

$O_A - O_B$	Concepts	Relations	Propriétés	Instances	T	Ress. Externe
Manuel.	13 - 13	15 - 14	-	11 - 0	24	Expert
OntAlign	9 - 9	12 - 11	3 - 3	11 - 0	-	WordNet

Les deux ontologies de cette paire sont très proches aussi bien au niveau structure qu'au niveau linguistique. La première a 11 instances tandis que la deuxième n'en a aucune, mais cela n'influence pas la performance de notre algorithme OntAlign. C'est ces éléments qui donnent à l'ontologie toute sa puissance en matière d'expressivité et permettent de réaliser des inférences.

Les deux ontologies, il est clair, modélisent le même domaine ; elles présentent des similarités évidentes. Cependant, elles présentent aussi des différences.

A titre d'exemples, nous remarquons que dans la première ontologie  $O_A$  le concept d'*hermaphrodite* n'a pas été évoqué, ce qui a été fait au niveau de la deuxième  $O_B$ . *hasMother*, *hasMom*, *hasFemalParent* sont utilisés comme équivalents dans  $O_A$ , alors que dans  $O_B$ , on s'est contenté du terme *hasMother*. Dans la première ontologie, on précise que *HumanBeing* veut dire la même chose que *Person*, tandis que dans la deuxième on ne le souligne pas. Pour décrire la même chose, les deux ontologies utilisent des termes différents : *TwoLeggedThing*, *TwoLeggedPerson* dans la première et *BipedalThing*, *BipedalPerson* dans la seconde. Ces différences suffisent pour empêcher une exploitation effectivement efficace des deux ontologies de manière conjointe. Il faudra donc les apparier.

La différence entre ces deux ontologies est la définition des concepts et des propriétés équivalentes. Par exemple, la première ontologie définit *hasFemaleParent* comme la propriété équivalente de la propriété *hasMother*, et *HumanBeing* comme la classe équivalente de la classe *Person*. L'application OntAlign exécute un post-traitement permettant de détecter des entités équivalentes. Cela permet de trouver toutes les correspondances de ce type.

**Tabl. 2.** Résultat de OntAlign sur la paire d'ontologies « Animals »

Seuil	F	T	C	P=C/F	RI=C/T	f-mesure	Overall
0.70	32	24	32	1.000	1.333	1.142	1.333
0.80	26	24	26	1.000	1.083	1.039	1.083
0.82	24	24	24	1.000	1.000	1.000	1.000
0.85	18	24	18	1.000	0.750	0.545	0.750
0.90	18	24	18	1.000	0.750	0.545	0.750
1.00	6	24	6	1.000	0.250	0.400	0.250

Avec le seuil de similarité de 0,82 pour OntAlign, le *f-mesure*=1.000 et toutes les correspondances correctes sont bien trouvées (1.000, tabl. 2). Si nous augmentons ces seuils, le nombre des correspondances correctes mais non-détectées est aussi augmenté, et donc, la valeur de *f-mesure* diminue.

##### ♦ La paire « Computer Networks »

Les deux ontologies suivantes ont une structure similaire et les termes nommant les entités le sont aussi. Par contre, la structure ou la taxonomie d'ontologie est assez complexe, avec la profondeur maximale de 4.

**Tabl. 3.** Informations concernant la paire d'ontologies « Computer Networks »

O <sub>A</sub> – O <sub>B</sub>	Concepts	Relations	Propriétés	Instances	T	Ress. Externe
Manuel.	27 – 27	5 - 6	-	0 - 0	30	Expert
OntAlign	27 - 27	5 - 6	0 - 0	0 - 2	-	WordNet

L'algorithme OntAlign se base principalement sur la linguistique, il détecte incorrectement donc que la classe *FTPServer* de la première ontologie, qui est sous-classe de *ServerSoftware*, correspond à la classe *Server* de la deuxième ontologie, qui est sous-classe de *Computer*, car la valeur de similarité totale calculée est de 0,80 par Wu & Palmer[8] et de 0,90 par la mesure terminologique (basé string). Tandis que la valeur de similarité entre *FTPServer* et sa vraie classe correspondante dans la deuxième ontologie, *FTP*, n'est que 0,50 par Wu & Palmer[8] et 0,825 par la mesure terminologique *StringDistance*. L'algorithme OntAlign prend en compte la structure d'ontologie et trouve donc correctement la classe correspondante de la classe *FTPServer*, qui est la classe *FTP* (sous-classe de *ServerSoftware*) à 0,94 par *StringDistance*.

La valeur de f-mesure est maximale entre (1.016, 0.775 et 0.666) avec les seuils de similarité de 0.80, 0.825, 0.925 et 0.9 dans OntAlign, respectivement (tabl. 4). La valeur de *f-mesure* pour OntAlign est de 1.016  $\approx$  1.00, elle est la meilleure valeur pour ce test. Si nous diminuons les seuils, le nombre de correspondances incorrectes augmente et diminue dans le cas contraire.

**Tabl. 4.** Resultat de OntAlign sur la paire d'ontologies « Computer Networks »

Seuil	F	T	C	P=C/F	R=C/T	F mesure	Overall
0.70	76	30	76	1.000	2.533	1.433	2.533
0.80	31	30	31	1.000	1.033	1.016	1.033
0.852	19	30	19	1.000	0.633	0.775	0.633
0.90	15	30	15	1.000	0.500	0.666	0.500
0.925	15	30	15	1.000	0.500	0.666	0.500
0.951	11	30	11	1.000	0.366	0.535	0.366

Pour les deux paires d'ontologies, les résultats renvoyés par f-mesure reflètent la précision de l'algorithme. Les résultats sont donc bons.

Suite à ces résultats, et à d'autres expérimentations et évaluations antérieures sur l'ensemble de paires d'ontologies de test fournies dans le cadre de la campagne de tests I3CON, tous les résultats obtenus par OntAlign sont encourageants.

Cependant, les mesures de performance de notre stratégie via les mesures telles que la précision, le *rappel*, la *f-mesure*, l'*overall* montrent que les résultats renvoyés par notre OntAlign proposé sont de qualité.

## 4 Conclusion

Notre approche est donc une des approches attaquant le problème d'appariement d'ontologies et spécialement pour les ontologies représentées en OWL. Elle emploie les mesures de similarité basant sur la comparaison des chaînes de caractères inspirée du domaine de la recherche d'information et l'intégration de WordNet pour produire la similarité linguistique lexicale entre deux entités. Elle exploite aussi des heuristiques pour le calcul de la similarité de deux entités. Ainsi, notre algorithme est une application et une intégration des mesures de similarité de base [8], [11] en exploitant des caractéristiques du formalisme OWL et diffère des autres approches dans le domaine de l'appariement des schémas ou des ontologies.

Cependant des perspectives de notre travail s'orientent donc vers :

- Faire étendre le processus d'appariement aux restrictions, fonctions et axiomes,
- Explorer les mesures de similarité basées sur le contenu informationnel,
- Tester l'approche sur d'autres domaines d'application, tel Peer-to-Peer
- Interroger une autre ressource linguistique comme le Euro-Wordnet,...
- etc.

## Références

1. Deliverable : Stefano Spaccapietra coordinator (Ecole Polytechnique Fédérale de Lausanne) "D2.1.3.1 Report on Modularization of Ontologies", 2005
2. A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, A. Halevy, "Learning to match ontologies on the Semantic Web", *The VLDB Journal*, 12:303-319, 2003.
3. M. Ehrig, S. Staab, "QOM – Quick Ontology mapping", in proc. of the 3rd International Semantic Web Conference (ISWC2004), November 7 to 11, Hiroshima, Japan, 2004.
4. J. Euzenat, P. Valtchev, "An integrative proximity measure for ontology alignment", in Proc. ISWC 2003.
5. A. Maedche, S. Staab, "Measuring similarity between Ontologies", in proc. of the European Conference on Knowledge Acquisition and management – EKAW-2002, Madrid, Spain, October 1-4, LNCS/LNAI 2473, Springer, 2002, pp. 251-263, 2002.
6. G.A. Miller, "Word Net: A lexical Database for English", *Communications of the ACM*, Vol. 38, N°11, p. 39-45, November, 1995.
7. N. Noy, M. Musen, "Anchor-PROMPT: Using Non-Local Context for Semantic Matching", *IJCAI* 2001.
8. Wu Z. & Palmer M. Verb Semantics and Lexical Selection, *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133-138, 1994.
9. Noy, N. F. & Musen, M. A., PROMPT: Algorithm and tool for automated ontology merging and alignment, in 'Proc. 17th Natl. Conf. On Artificial Intelligence (AAAI2000)', Austin, Texas, USA 2000b.
10. Doan, A., Madhavan, J., Domingos, P. & Halevy, A., Ontology matching: A machine learning approach, in S. Staab & R. Studer, eds, 'Handbook on Ontologies in Information Systems', Springer-Verlag, pp. 397–416, 2004.
11. Zargayouna H. , Contexte et sémantique pour une indexation de documents semi-structurés. ACM Conférence en Recherche Information et Applications, CORIA'2004.
12. Kefi, H.,B. Safar, C. Reynaud, Alignement de taxonomies pour l'interrogation desources d'information hétérogènes. RFIA, Tours, 2006.
13. Jerome Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.
14. Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: a versatile graph matching algorithm. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 117– 128, 2002.

## *Software engineering I*

---



## Méta-fédération : une architecture pour l'assemblage d'éléments logiciels

Khaled Sellami<sup>1</sup>, Mohamed Ahmed-Nacer<sup>2</sup>

<sup>1</sup> Université de Bejaia, Route Targa-Ouzemour 06000, Bejaia , Algérie

[skhaled36@yahoo.fr](mailto:skhaled36@yahoo.fr)

<sup>2</sup> Laboratoire des Systèmes Informatiques - L.S.I. U.S.T.H.B.

B.P. 32 EL-ALIA, 16 111, BabEzouar, Alger, Algérie,

[anacer@mail.cerist.dz](mailto:anacer@mail.cerist.dz)

**Résumé.** Les fédérations de procédés logiciels sont un ensemble de composants logiciels mis à disposition sur l'Internet pour assister les développeurs de logiciels. Cependant les services offerts par ces fédérations sont peu exploités pour la construction de nouvelles applications logicielles. Dans ce contexte, nous proposons une architecture permettant l'assemblage de fédérations de procédés logiciels dans laquelle un groupe de fédérations de procédés logiciels est amené à travailler et à coopérer pour créer ces applications logicielles. Nous utilisons le concept de "fédération" comme une architecture logicielle qui matérialise cet assemblage, en définissant trois fondations (fondation de l'application, fondation de contrôle, fondation du procédé) comme structure de base d'une méta-fédération et cela comme solution aux problèmes de recouvrements des concepts des fédérations de procédés, d'indéterminisme de celles-ci, et de leurs adaptations.

### 1 Introduction.

Les fédérations de procédés logiciels sont largement utilisées pour administrer l'industrie du logiciel [1] [6]. Ces fédérations peuvent être vues comme des composants logiciels offrant des services sur le web [4]. Assembler ces composants pour la construction de logiciels de grande taille reste cependant trop complexe.

Dans cet article, nous présentons une architecture pour la construction de logiciels de grande taille. Cette architecture utilise les services offerts par ces composants « fédérations » en utilisant le mécanisme d'interopérabilité: nous parlons alors de méta-fédération.

Dans ce contexte, trois problèmes majeurs doivent être résolus [2] [3] :

1. le recouvrement (partiel) des fonctionnalités et des concepts des fédérations de procédés ou d'outils<sup>1</sup> de la méta-fédération,
2. les incohérences entre les concepts tels qu'ils sont définis et leurs implémentations au niveau des fédérations d'outils,
3. l'indéterminisme des comportements de chacune des fédérations de procédés ou d'outils dans le cadre d'une méta-fédération.

En effet, ces fédérations sont conçues pour être autonomes (interface propre, langages différents, plates formes différentes d'exécution, multiples moyens de communication,...).

Par conséquent, la méta-fédération doit offrir un moyen de description de la synchronisation entre le procédé abstrait, et l'ensemble des fédérations concrètes. Ainsi le procédé abstrait peut être vu comme une spécification exécutable, et les fédérations concrètes comme une implémentation de cette spécification (Fig. 1).

La section 2 présente l'architecture de notre méta-fédération

---

<sup>1</sup> . Dans la suite de l'article, nous utiliserons indifféremment le terme fédération de procédé et fédération d'outils car c'est un cas particulier des fédérations de procédés.

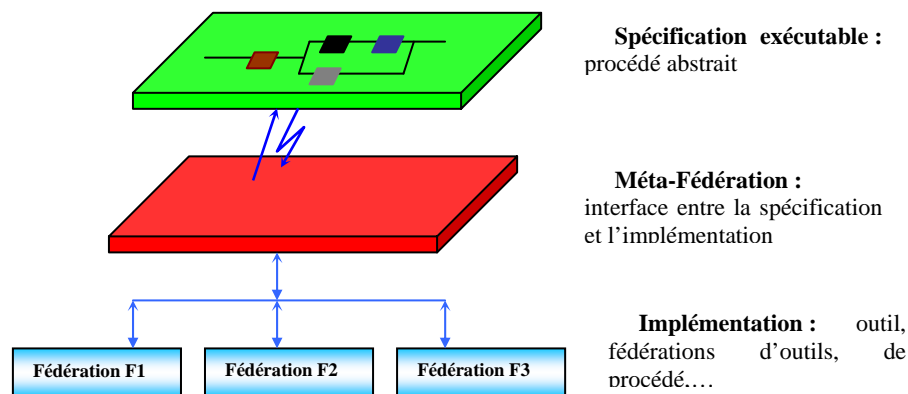


Fig. 1. Les différents niveaux d'abstraction de la méta-fédération

## 2 Architecture de la Méta-fédération.

Afin de faciliter la présentation de l'architecture de la méta-fédération, nous utilisons l'exemple de transfert d'un produit (ce produit étant un document de spécifications) d'une activité<sup>2</sup> de *spécification* vers une activité de *revue*. Ainsi, les fédérations de procédés sont impliquées dans différentes "sections", "aspects" du document dans le cas du scénario de transfert. Pour créer cette méta-fédération, on aura besoin des services de plusieurs composants des différentes fédérations : l'outil CVS (concurrent Version Control) pour gérer les différentes versions du document, un outil de traitement de texte, quelques outils d'agenda sur lesquels les développeurs travaillent.

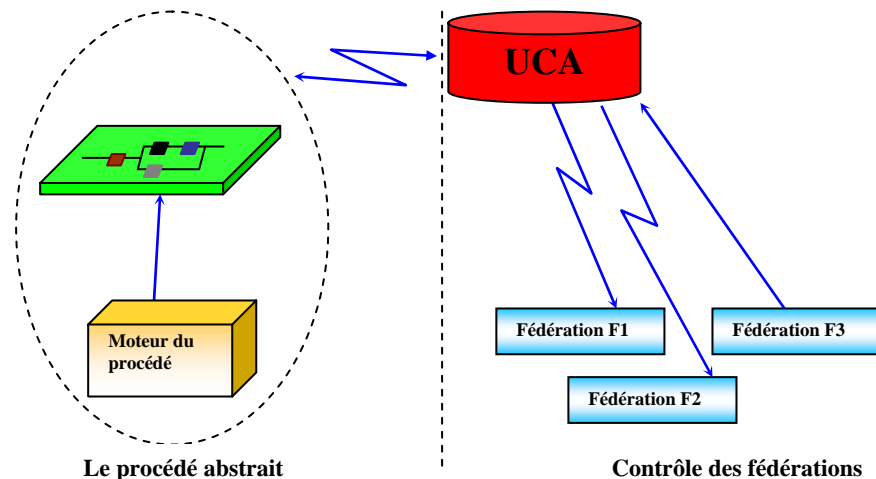
### 2.1 L'univers Commun.

En effet chacune de ces fédérations doit "connaître" le document à transférer, mais chacune décline sa propre "vision" du document en fonction de son univers de discours. Nous proposons de définir un ensemble de concepts communs et selon un haut niveau d'abstraction, permettant de mettre sous un même édifice l'ensemble des connaissances communes et des concepts manipulés par les fédérations composant la méta-fédération. Ces concepts "communs" sont directement liés à l'application de la méta-fédération. Cela se traduit, au niveau de l'architecture d'une fondation par un composant appelé "univers commun (UC)" permettant le stockage des entités et de leurs concepts, une interface (API) permettant la manipulation de ces entités (consultation et modification de l'état de l'univers commun) ainsi qu'un gestionnaire de notifications. Les services de la fondation (l'UC, les fédérations participantes, et les mécanismes d'interactions entre l'UC et ces fédérations) sont appelés les services de la méta-fédération ou, services fédéraux. Selon ce modèle, la connaissance "commune" à l'ensemble des fédérations d'outils et permettant leur synchronisation est "librement" gérée. Il n'y a aucun contrôle de cohérence de la méta-fédération (absence de règle de fonctionnement explicite).

Le but de la méta-fédération est la réalisation de l'application. Cette réalisation dépend des comportements du concepteur (humain) qui sont souvent imprévisibles utilisant des technologies qui lui sont mises à disposition. Il revient alors de résoudre ce problème d'indéterminisme dans la méta-fédération, et du fait qu'un modèle de procédé décrit le comportement attendu du monde réel et l'évolution attendue de ce dernier. Nous proposons de définir un modèle de procédé qui décrit le comportement attendu de l'univers commun de l'application (UCA) et l'évolution attendue de ce dernier. Ce modèle ainsi défini peut être vu comme l'expression formelle du procédé permettant d'atteindre les buts de la méta-fédération, c'est-à-dire, permettant de réaliser l'application. Dans ce cas il y a des informations dans l'UCA qui sont liées au modèle de procédé. La figure (Fig. 2) ci-dessous montre le cas où les outils de procédé sont tous utilisés comme des fédérations participantes.

<sup>2</sup> . Une activité est une étape du procédé dans laquelle une action est exécutée.





**Fig. 2.** Evolution des fédérations d'outils par un modèle de procédé

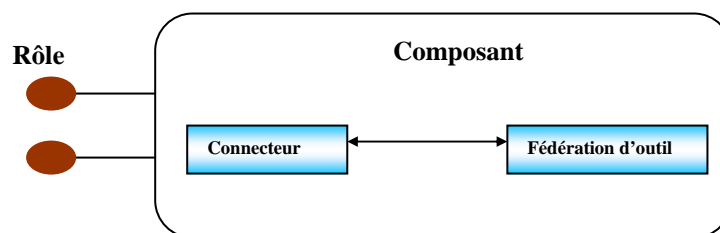
Ce modèle de procédé fait partie d'une deuxième fondation qu'on appelle fondation du procédé (qu'on a défini par analogie à la définition générale d'une fondation) qui est constituée d'un univers commun, appelé ici univers commun du procédé (UCP), contenant un modèle du procédé, Un ensemble d'outils impliqués dans cet univers commun de procédé (outils de monitoring, outils de stratégie d'évolution, outils de décisions pour la réconciliation, outil d'implémentation des changements : outils propre à la technologie du procédé), et mécanismes d'interactions permettant aux outils de la fondation de procédé (FP) de communiquer avec l'UCP.

## 2.2 Gestion de l'hétérogénéité.

Dans la méta-fédération, les fédérations d'outils peuvent offrir des services d'outils utilisant des concepts similaires. Cependant, la représentation de ces concepts dans l'UCA et les fédérations participantes peut être tout à fait différente, d'où la nécessité de faire correspondre l'ontologie de l'univers commun et l'ontologie de chacune des fédérations d'outils (et vice-versa); ainsi est-il nécessaire d'adapter les fédérations d'outils.

Nous proposons de définir un niveau d'adaptation entre les fédérations d'outils et l'UCA pour faire la correspondance entre les différents rôles définis dans le cadre de la méta-fédération et les fédérations d'outils qui réalisent ces rôles.

Dans ce niveau d'adaptation les composants utilisés sont appelés connecteurs (fig.3). Ces derniers effectuent les concordances syntaxiques et sémantiques nécessaires entre les concepts et services des fédérations d'outils, et les services déclarés des rôles endossés par les fédérations d'outils. Le connecteur remplit les fonctions principales suivantes : Effectuer les traductions entre l'UCA et la fédération d'outil et l'inverse. Et mettre en place et établir la communication avec la fédération d'outil dont il est le représentant en utilisant les mécanismes d'interaction de la méta-fédération. On appelle composant l'ensemble fédération d'outil et son adaptateur.



**Fig. 3.** Adaptation des fédérations d'outils participantes

La méta-fédération observe les fédérations d'outil via leurs rôles. Tous les contrôles qui se font au niveau des fédérations d'outils seront représentés dans la section suivante, et sont matérialisés par les rôles qui sont des composants abstraits. Ainsi, la fédération d'outil et le connecteur représente le composant concret qui satisfait les propriétés du composant abstrait. La séparation du composant abstrait et concret nous permet de recenser les services similaires offerts par les différentes fédérations d'outils.

### **2.3 Le contrôle dans la méta-fédération.**

En reprenant le scénario issu du domaine des procédés logiciels, nous constatons que les fédérations d'outils de notre méta-fédération présentent des similitudes à la fois dans certains concepts comme nous l'avons vu mais également dans leurs fonctionnalités. Il est possible que plusieurs fédérations d'outils fassent la même chose simultanément (par exemple, l'édition d'un même fichier – et donc deux services d'édition de fichiers différents.). Ceci implique une modification d'une même entité de l'univers commun simultanément ; ce qui a pour conséquence de produire une incohérence au niveau de l'état de ce dernier. La "libre entreprise" et l'initiative laissée aux outils sont source d'incohérence de l'univers commun. Il serait donc intéressant de pouvoir disposer de mécanismes permettant de contrôler les initiatives des fédérations d'outils, en limitant par exemple leurs possibilités, en dédiant un "aspect" de l'UCA à une fédération d'outils spécifique, etc. Il revient alors de contrôler : Les actions des fédérations d'outils sur l'UC, Les réactions potentielles des fédérations d'outils face aux changements d'état de l'UC, on ajoute aussi le contrôle de la coordination des actions des fédérations d'outils.

Nous proposons de placer un niveau de contrôle entre les fédérations d'outils et l'UCA qui utilise un univers commun de contrôle (UCC) (Figure 4). Cet univers commun permet de stocker les informations concernant le contrôle des initiatives, d'observation et de coordination entre les fédérations d'outils.

#### ***Contrôle d'observation.***

Cette partie de contrôle concerne les notifications permises (le fait de recevoir des notifications lorsque le service fédéral en question a été invoqué)

#### ***Contrôle des initiatives.***

C'est-à-dire, de contrôler les initiatives des fédérations d'outils dans le cadre de la méta-fédération. Ainsi, le concepteur de la méta-fédération va pouvoir définir des interdictions concernant les fédérations d'outils. Une interdiction associe un service fédéral et une fédération d'outils en qualifiant s'il s'agit d'une invocation (le fait d'invoquer le service fédéral en question)

Le but est de donner la possibilité, au concepteur de la fédération, de restreindre le comportement des fédérations d'outils de la méta-fédération en inhibant tout ou partie de leurs initiatives (inhibition des "libertés").

#### ***Contrôle de cohérence et de coordination.***

Le comportement défini par le concepteur de la méta-fédération (transaction, ordonnancement des actions des fédérations d'outils, etc.) sera exécuté par ce niveau de contrôle par invocation des fédérations d'outils, et cela en interprétant un modèle de coordination qui aura été stocké dans l'UCC, donc dans ce dernier on a trois modèles : un modèle de contrôle des initiatives individuelles, un modèle décrivant les fédérations d'outils, et un modèle de coordination.

D'où une fondation de contrôle (constitué de l'UCC, les fédérations d'outils participantes, et les mécanismes interactions) dont le rôle est justement d'intercepter les requêtes en destination de l'UCA (seul connu des fédérations d'outils) et d'exécuter le modèle de coordination à l'insu des fédérations d'outils. Et dans ce cas là les fédérations d'outil n'ont plus aucune liberté (solicitation synchrone, point à point).

Pour matérialiser le niveau de contrôle de coordination, nous introduisons le concept d'aspect : Chaque aspect contient l'expression de la coordination des différentes fédérations d'outils, exécutée lorsqu'un service fédéral est invoqué. D'où l'utilisation de la technique de POA (la programmation orienté aspect) [5] pour modifier et supprimer dynamiquement les aspects de notre système.

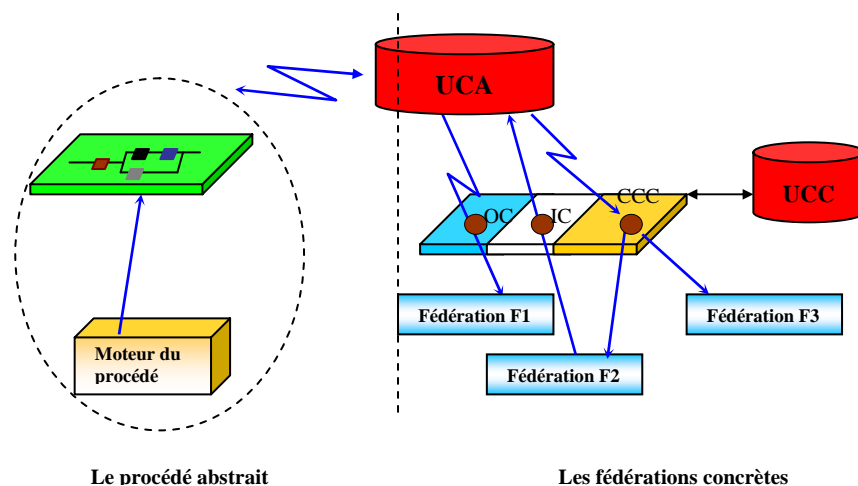


Fig. 4. Contrôle d'observation, de cohérence, et des initiatives de la méta-fédération

## 2.4 Développement.

Notre méta-fédération est en phase de développement pour expérimentation (recensement de l'ensemble des concepts pertinents de l'application, expression des concepts à l'aide d'un formalisme spécifique,...). L'application étant distribuée sur le réseau, on utilise l'appel de procédure à distance (RMI ou les Sockets pour assurer la communication entre le Wrapper et le Proxy qui représentent le connecteur). Nos efforts vont s'axer sur le développement et l'administration de la méta-fédération, à travers la définition d'une interface graphique pour désigner de manière simple et rapide notre méta-fédération.

## 3 Conclusion.

Nous avons présenté dans cet article nos travaux sur le problème des fédérations de composants logiciels hétérogènes, autonomes, et distribuées sur le réseau dont l'objectif est de proposer une approche permettant la conception de ces fédérations en tenant compte de toutes ces contraintes.

L'architecture que nous proposons s'inspire des travaux de [2] et [3] sur les fédérations d'outils qu'on a repris pour une fédération de fédérations d'outils logiciels, qu'on a appelé méta-fédération. Cette méta-fédération repose sur trois fondations : fondation de l'application, fondation du procédé, et fondation de contrôle.

Nous pensons que nos travaux facilitent les tâches incombant au concepteur de ces systèmes par le découplage des différentes tâches; ils se concentrent prioritairement sur la phase de conception, facilitant ainsi les phases de codage et d'intégration.

## 4 Bibliographie.

1. M. Ahmed-Nacer, M-A. Mostefai: Solving Interoperability Problems on a Federation of Software Process Systems. (ICEIS (1)) : 591-594, Porto, Portugal 2004
2. J. Estublier, H. Verjus, P.Y. Cunin, "Building Software Federation", Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'2001), Las-Vegas, EU, juin 2001.
3. J. Estublier, H. Verjus, P.Y. Cunin, "Modelling and Managing Software Federation", Proceedings of the European Conference on Component-Based System Engineering (CBSE'2001),
4. J. Estublier, S. Sanlaville, "Mélusine: un environnement de modélisation et de coordination de services", Revue d'Ingénierie des Systèmes d'Information, numéro spécial "Services Web, théories et applications", Volume 10, Numéro 3/2005
5. K. Lieberherr, D. Orleans, J. Ovlinger, "Aspect-Oriented Programming with Adaptive Methods", Communications of the ACM, Vol. 44, No. 10, pp. 39-41, October 2001.
6. S. Bandinelli, M. Barga, A. Fuggetta, C. Ghezzi, L. Lavazza, "SPADE: An Environment for Software Process Analysis, Design and Enactment", Software Process Modeling and Technology, Research Studies Press, Taunton, 1994.

# Structural and Behavioural Composition Aspects in Component-Based Architectures

Abdelkrim Amirat and Mourad Oussalah

Laboratoire LINA CNRS FRE 2729, Université de Nantes  
2, Rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03, France  
{Abdelkrim.Amirat, Mourad.Oussalah} @Univ-Nantes.fr

**Abstract.** A Critical level of abstraction in the description of a complex system is its software architecture. At an architectural level, one describes the principal system components and their pathways of interaction. Architecture is considered to be the driving aspect of the development process; it allows specifying which aspects and models in each level needed according to the software architecture design. Early Architecture Description Languages (ADLs), nearly exclusive, focus on structural abstraction hierarchy ignoring behavioural description hierarchy, conceptual hierarchy, and metamodeling hierarchy. In this paper we show that all those hierarchies constitute views to appropriately “reason about” software architectures described using our C3 metamodel which is a minimal and complete ADL. We provide a set of mechanisms to deal with different levels of each hierarchy; also we introduce a new enhanced definition for connector concept deployed in C3 architectures.

**Keywords:** Metamodel, Abstraction, Behaviour, Conceptual component, Connector, Composition, Service.

## 1 Introduction

Software architectures are composed of components, connectors and configurations, constraints on the arrangement and behaviour of components and connectors. The architecture of a software system is a model, or abstraction of that system. Software architecture researchers need extensible, flexible architecture descriptions languages (ADLs) and equally clear and flexible mechanisms to manipulate these core elements of the architecture. Recently Medvidovic [4] gives the following definition for software architecture “*A software system’s architecture is the set of design decisions about the system*”. So, if those decisions are made incorrectly, they may cause your project to be cancelled. However, these design decisions encompass every aspect of the system under development, including:

- Design decisions related to system *structure*;
- Design decisions related to behaviour also referred to as functional;
- Design decisions related to the system’s non functional properties;
- Also, we can elicit other design decisions related to the development process or the business position (*product-line*).

However, we note that in the ADLs which currently exist, there is no standard about architectural concepts or standards in terms of mechanisms for manipulating those concepts. For the reasoning model, the majority of ADLs proposes only sub-typing as a mechanism for specialization (e.g. Acme, C2 [2]). Otherwise, for the rest of ADLs, they propose their own ad hoc mechanisms based on algorithms and methods designed specially for these ADLs. Furthermore, all ADLs capture aspects of software design centred around a system’s *component*, *connectors*, and *configurations*. So, the core elements of our model are basically defined around these tree elements. We derive the name of our model C3 for Component, Connector, and Configuration. The rest of the paper is organized as fellows. In section 2 presents our research motivations. Section 3 describes the C3 metamodel. A case study is presented in section 4. The last section presents our conclusion and the future perspectives.

## 2 Motivation

Describing architecture’s structure is not sufficient to provide correct and reliable software systems. In this paper we are even more going to focus on representation architecture model and how to reason about

its elements following four different types of hierarchies. Each of these hierarchies provides a particular view on the architecture. In the following sections we present more details about these hierarchies. We expect from our approach to provide a more explicit and better clarified software architecture description. Mainly the approach is developed to:

- Make explicit the possible types of hierarchies being used as support of reasoning on the architecture structures, with the different possible levels in each hierarchy.
- Show semantics convey by every type of hierarchy by providing the necessary mechanisms used to connect elements.
- Allow introducing various mechanisms of reasoning within the same architecture according to the requirements of the system in a specific application domain.
- Establish the position of existing mechanisms developed for reasoning with regard to our referential.

### 3 The Metamodel of C3

In order to have a complete C3 model, we have defined mainly two complementary models to describe and reason about system's architecture. We use *representation model* to describe architectures based on C3 elements and we use *reasoning model* to understand and analyse the representation model.

#### 3.1 Representation Model

The core elements of the C3 representation model are components, connectors, and configurations, each of these elements have an interface to interact with its environment. Fig. 1 depicts the core elements of the metamodel C3.

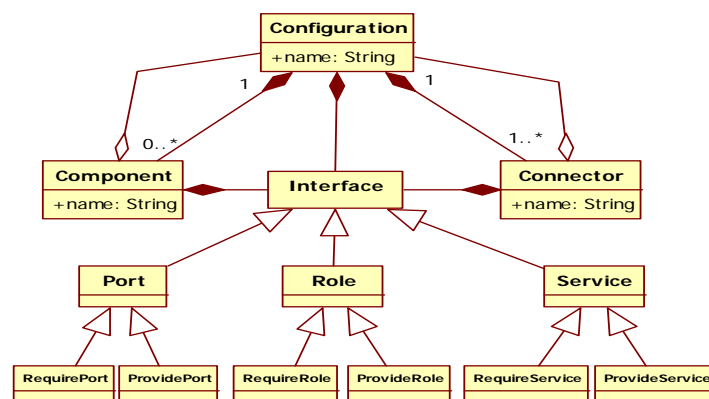
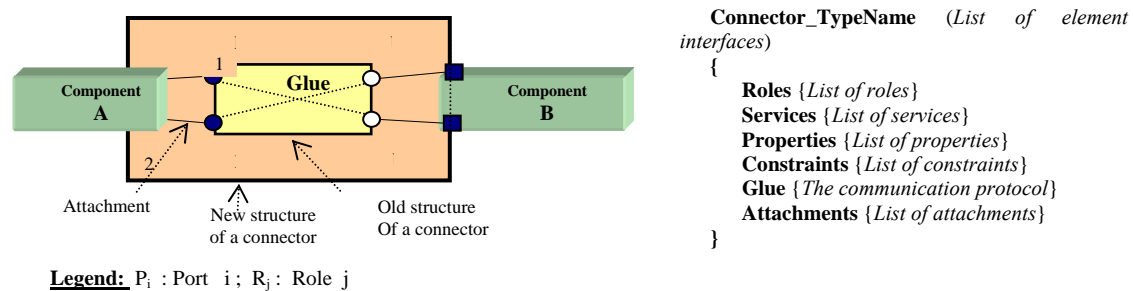


Fig. 1. Basic elements of C3 Meta model

**Components** represent the primary computational elements and data stores of a software system. Intuitively, they correspond to the boxes in box-and-line descriptions of software architecture. In C3, each component can have one or more ports. Ports are the interaction points between components and their environments.

**Connectors** in C3 represent interconnections among components for supporting their interactions and they are defined explicitly and considered as first class entities by separating their interfaces (roles) from their services (glues). Fig. 2.a illustrates our contribution at this level which consists in enhancing the structure connectors by encapsulating the attachment links inside the connector. So, the application builder will have to spend no effort in connecting connectors with its compatible components/configuration. Consequently, the task of the developer consists only in selecting a suitable type of connector which is compatible with the types of components/configurations which are expected to be connected. We have given the following syntactic definition for connectors in a previous work (Fig. 2.b) [1].



**Configurations.** In C3 representations model each component or connector is perceived and handled from the outside as primitive element. But their inside can be real primitive elements, or composite with a configuration to encapsulates all the internal elements [8]. These Configurations are first-class entities. A configuration may have ports, and each port is perceived like a bridge between the internal environment of the configuration and the external one. In general, configurations can be hierarchical where the internal components and connectors can represent sub-configurations with their proper internal architectures and so on.

### 3.2 Reasoning Model

In our approach software architecture is analysed using different hierarchy views where each hierarchy is investigated at different levels of representations. This model is defined by four types of hierarchies: 1- The structural hierarchy used to explicit the different nested levels of structural hierarchies that the system's architecture can have. 2- The behavioural description hierarchy to explicit the different levels of system's behaviour hierarchy generally represented by protocols. 3- The conceptual hierarchy to describe the libraries of element types corresponding to structural or behavioural element at each level of the architecture description. 4- The metamodeling hierarchy to locate where our model coming from and what we can do with it. Obviously those two sides will belong to the pyramid of hierarchies defined by OMG [5].

**Structural hierarchy (SH).** Also called abstraction hierarchy, has to provide the structure of particular system architecture in terms of the architectural elements defined by the used ADL. The majority of academic ADLs like Aesop, MetaH, Rapide, SADL, and other [3] or the industrials like CORBA, CCM/CORBA, EJB/J2EE [6] allow only a flat description of software architectures. Using those ADLs architecture is described only in terms of components connected by connectors without any nested elements - without any structural hierarchy. This design choice was made in order to simplify the structure and also by lack of concepts and mechanisms that respectively define and manipulate configurations of components and connectors.

In our C3 model the structure of architectures is described using components, connectors, and configurations, where configurations are composite elements. Each element in this configuration (component or connector) can be a primitive or a new configuration which contains another set of components and connectors, which in their turn can be primitive or composite material, and so on.

However, the metamodel C3 allows the representation of architecture with a real hierarchy with a number of abstraction levels depending on the complexity of the problem. It is important to note that practically all architectural solutions for domain problems have a nested hierarchical nature. Thus, real software architecture can be viewed as a graph where each internal node of this graph represents a configuration and each end-node represents a primitive component arcs between nodes are connectors. In Fig. 3.a, the root node with double circles represents is the first level of the abstraction; it is also the global configuration which encapsulates all elements of the architecture. The small white circles represent primitive components and small black circles represent sub-configurations in the system architecture. These configurations contain other elements inside. Thus, a configuration will never be an end-node in the hierarchy tree of abstractions. Arcs represent the bonds of levels - the father/child relationship. This relationship does not necessarily imply a service-connection between the father node and the child one. To explicit the composition relation among elements we define the following type of connector:

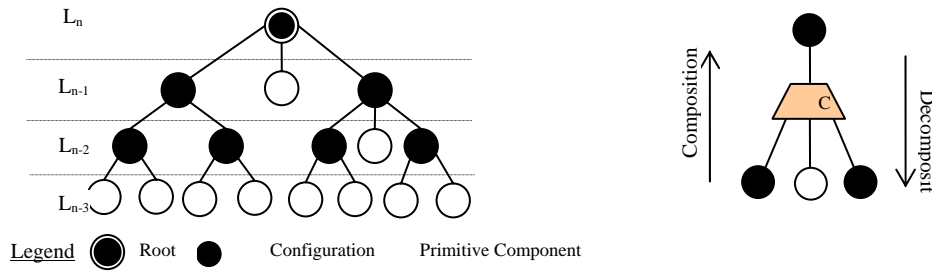


Fig. 3. a. External view of the SH (1).

b. CDC connector

*Composition-Decomposition Connector (CDC)* used to link each configuration to its underlying elements. Therefore, this type of connector allows the navigation among levels of the structural hierarchy. So, we can determine the children or the father, if it is the case, of each element deployed in the architecture. Fig. 3.b illustrates how to use CDC connector to represent abstraction details (nodes 7, 8, and 9) at level  $n-2$  corresponding to node (4) at level  $n-1$ . So, the internal view of Fig. 3.a is described by seven CDC connectors.

*Attachment Connector (AC)* we use this type of connectors to establish service-connections between components and/or configurations deployed in the same level of abstraction (Fig. 4.a). In some ADLs this type of connector is called *assembling connector* and represented by first class entity (e.g. Acme [2]). Inside the AC connector the glue code defines the mapping among communicating elements.

*Expansion-Compression Connector (ECC)* is represented in Fig. 5.b. is used to establish service-connections between each configuration and its underlying elements (Fig. 4.b). In some ADLs this type of link is called *binding or delegation* but it is not defined as a first class entity.

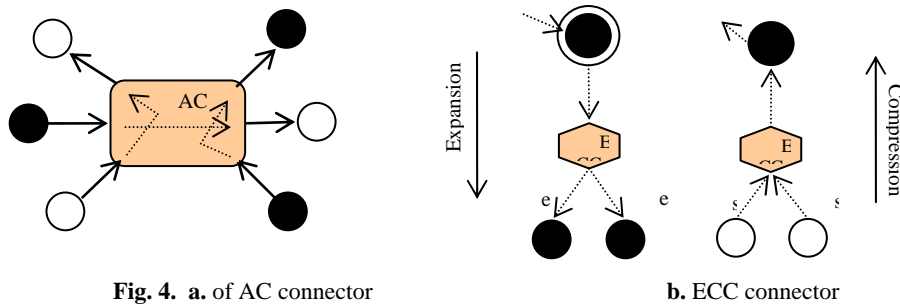


Fig. 4. a. of AC connector

b. ECC connector

The different elements of the architecture are connected through their interfaces. Thus the types of interfaces are checked if they are compatible or not. Consequently, in the structural hierarchy, the consistency of elements assembly is controlled syntactically.

**Behavioural hierarchy (BH).** The behavioural hierarchy represents the description of the system's behaviour at different levels. Each primitive element of the architecture has its own behaviour. The behaviour description associated with the highest level of the hierarchy - level  $n$  - in Fig. 5 represents the overall behaviour of the architecture. This behaviour is described by a global protocol  $P_0$ . The system architecture at this level is perceived as a black box with a set of inputs (*required services*) and outputs (*provided services*). At lower level each component, connector, configuration, port, or role has its own protocol to describe its functionality (e.g. glue code is the protocol describing the connector behaviour, also the component behaviour can be described by a statechart diagram). So, a protocol is a mechanism used to specify the behaviour function of an architectural element by defining the relationship among the possible states of this element and its ability to produce coherent results. Fig. 6.a sketches how to decompose the protocol  $P_0$  at level  $n$  into its sub-protocols at level  $n-1$ . This decomposition process produces a set of other protocols  $\{P_{01}, P_{02}, P_{03}\}$ . By the same process each protocol of the level  $n-1$  is decomposed to produce another set of sub-protocols at the level  $n-2$ , and so on until level  $0$ . The last level of the hierarchy is a set of protocols representing the primitive behaviour of elements which are available in the

library of the architect. The total set of protocol levels represents the behaviour hierarchy of the system architecture.

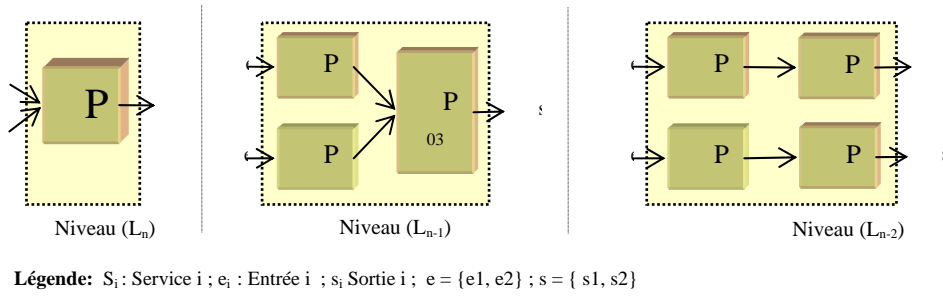


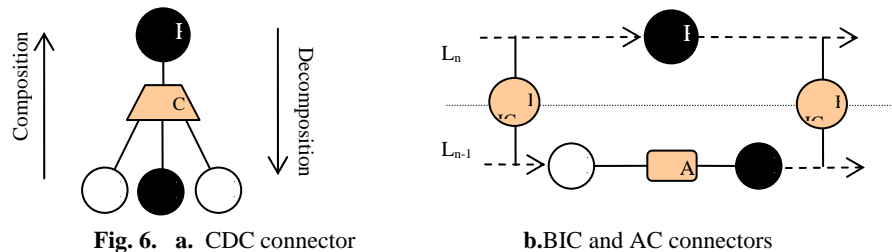
Fig. 5. External view of BH

To explicit the parental relationship among elements belonging to successive levels of the behavioural hierarchy we define the following types of connector:

*Composition-Decomposition Connector (CDC)* used to link each protocol to its possible sub-protocols. Thus, this type of connector allows the navigation among levels of the behavioural hierarchy. Also, we can determine the childs or the father, if it is the case, of each protocol used in the architecture. Fig. 6.a represents the notation adopted.

*Attachment connector (AC)* used to connect protocols belonging to the same level of hierarchy. If we use, for example, transition-based system to specify the behaviour protocol associated with each element then connections between behaviours are made by simple transitions between the end-state of the first protocol and the start-state of the second one. Inputs and outputs of each protocol are respectively required and provided services (Fig. 6.b).

*Binding Identity Connector (BIC)* used to keep the identity and the traceability of inputs and outputs of protocols. In contrast to the structural hierarchy, in the behavioural hierarchy we have neither expansion of inputs nor compression of outputs when we change the level of hierarchy. Consequently, we use the same set of mechanisms and tools at all hierarchical levels. So, when we move to the below level this in fact amounts to just slicing the behaviour function to several sub-functions in order to understand the global complex behaviour of the system (Fig. 6.b).



**Conceptual hierarchy (CH).** The conceptual hierarchy allows the architect to model the relationship among elements of the same family as illustrated in Fig. 7.a for the component types. The architectural entities are represented by types. Each type is an element in the library and each element has its sub-elements in the same library. So, we can shape the graph representing entities hierarchy of the same family. Each graph has its proper number of levels (sub-type levels). At the highest level of hierarchy we have the basic element types developed to be reused. The element types of the intermediate levels are created by reusing the previous ones. Those intermediate element types are reused to produce others or used as end-elements to describe architectures, and so on. Element types at the last level of the hierarchy are only created to be used in the description of architectures [8]. Through the mechanism of specialization the architect will create and classify element libraries according to architecture development needs in each target domain. The number of sub-type levels is unlimited. But we must remain at reasonable levels of specialization in order to keep compromise between the use and the reuse



of the architectural elements. To implement the conceptual hierarchy we define the following type of connector:

*Specialisation-Generalisation Connector (SGC)* is used to connect element types coming from the same type. So, we can construct easily all trees representing the classification library types. Fig. 7.b represents the notation adopted.

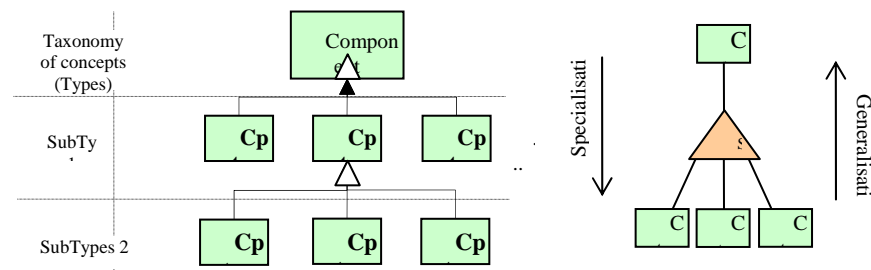


Fig. 7. a. External view of the conceptual hierarchy

b. SGC connector

**Metamodeling hierarchy (MH).** The metamodeling hierarchy is viewed like pyramid composed exactly from 4 architectural levels endowed with an instantiation mechanism. Thus, according to Fig. 8.a, each level ( $A_i$ ) must conform to the description given above in  $A_{(i+1)}$  level. The level  $A_3$  conforms to itself. Symmetrically, each level ( $A_i$ ) describes the inferior level  $A_{(i-1)}$ .  $A_0$  is the *application level* [1].

**$A_0$  Level** is the real word level (*application level*) which is an instance of the architecture model (level  $A_1$ ). At this level the developer has the possibility to select and instantiate elements any times as he needs to describe a complete application. Instances are created from element types which are defined at  $A_1$ . Elements are created and assembled with respect to the different constraints defined at  $A_1$  level.

**$A_1$  Level** is also called *architecture level*. At this level we have models of architecture, possibly with a given style, described using language constructions or notations defined at  $A_2$  level (e.g. C3 metamodel, UML 2.0).

**$A_2$  Level** (*meta-architecture level*) defines the language or the notation used to describe architectures at  $A_1$  level. This level is also used to modify or adapt the description language. All operations undertaken at this level are always in conformance with the top level of the pyramid.

**$A_3$  Level** (*meta meta-architecture*) has the top level concepts and elements used when we want to define any new architecture description language or new notation. In our previous work we have defined our proper meta meta-architecture model called Meta Architecture Description Language (MADL) [7]. So, our C3 metamodel is defined in conformance with MADL. MADL is similar to MOF but it is a component-oriented. To connect each architectural element to its type at the above level we define the following type of connector:

*Instance-Of Connector (IOC)* is used to establish connection among elements of a given level (*model*) with their classifier defined in the above level (*metamodel*). Fig. 8.b represents an example of the notation adopted.

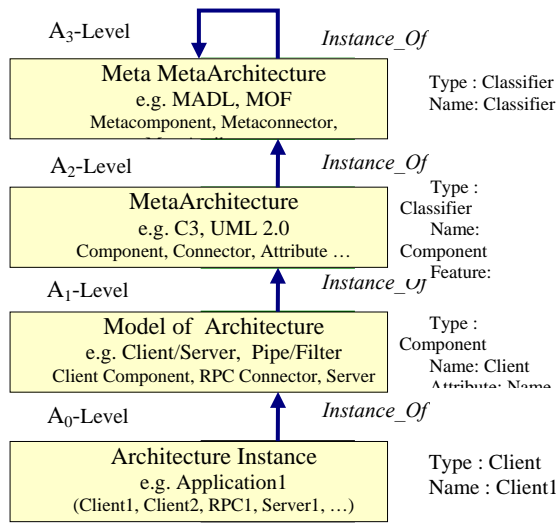
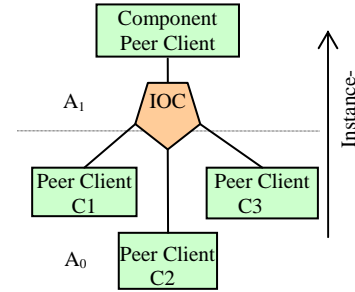


Fig. 8. a. Metamodeling hierarchy



b. IOC connector

#### 4 Case Study

As a simple illustrative example, Fig. 8 depicts: In the following subsections we try to analyse this example from the provided view of each type of hierarchy introduced in this paper. In the next Fig. we use numbers to represent the following elements: (1) Client-Server Architecture, (2) Client Component, (3) Server Configuration, (4) Connection Manager Component, (5) Security Manager Component, (6) Data Bases component. At level  $A_1$  we have the architecture model and at level  $A_0$  an example of application architecture

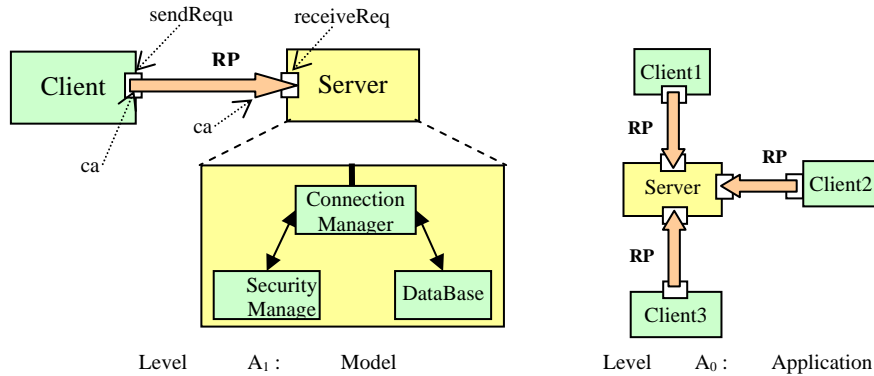
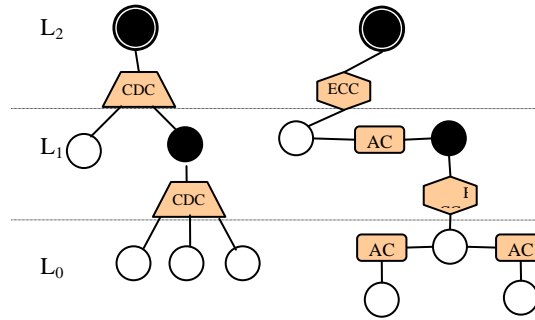


Fig. 8. Client-server model / application architecture

**Structural hierarchy** corresponding to this example can be represented by three levels. In Fig. 9 we represent this hierarchy by means of two graphs. Each graph gives particular view for the same set of nodes (six nodes). There is one node for each element of system. The left one illustrates the composition-decomposition view for the structural hierarchy using two CDC connectors, and the right one illustrates the same hierarchy but from physical structure interconnections view using two ECC connectors and four AC connectors. So, at this point we can say that all structural elements of the client-server system are depicted by the structural hierarchy.



**Fig. 9.** Structural hierarchy with CDC, ECC, and AC Connectors

In Fig. 10 we give, in an illustrative example, some details about AC connector representing the RPC connector used to connect the client component (node 2) to the server configuration (node 3).

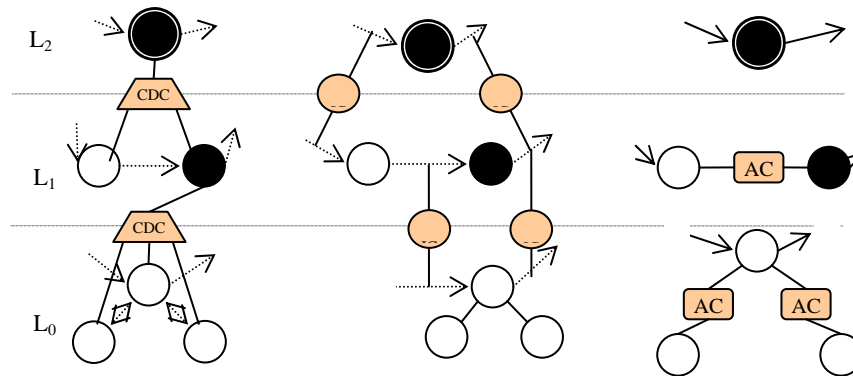
```

Component Client {Port {send-request}}
Configuration Server {Port {receive-request}}
Connector RPC: AC (Client, Server) {
  Roles {Caller, Callee}
  Services {List of services}
  Properties {maxRoles: integer = 2, Synchronous: Boolean = true}
  Constraints {List of constraints}
  Glue {caller = callee}
  Attachments {Client.send-request to RPC.caller, Server.receive-request to
RPC.callee} }

```

**Fig. 10** Description of RPC Connector

**Behavioural hierarchy.** is described by three graphs. Each graph illustrates the behavioural hierarchy using a particular type of connector. In Fig. 11.a we use two CDC connectors to describe the compression and decompression of the behaviour protocols. So, at level<sub>2</sub> we have the global protocol (P0), at level<sub>1</sub> we define (P1) client protocol and (P2) server protocol, and finally at level<sub>0</sub> we have (P4), (P5), and (P6) protocols associated respectively to connection-manager, security-manager and database components. In Fig. 11.b we represent exactly the same hierarchy in which we show the traceability of inputs and outputs using four BIC connectors. But in Fig. 11.c we only focus on continues flow of protocols at each level of the behavioural hierarchy using three AC connectors.



**Fig.11 a.** BH with CDC.

**b.** BH with BIC

**c.** BH with AC

**Conceptual hierarchy.** In Fig. 12.a we use the SGC connector to generate the five meta connector types from the first meta connector defined by C3. The SGC meta connector is the bootstrap connector for the others meta connector types, and by the same way we can use the SGC connector to specialise any architecture element described either at level A<sub>1</sub> or at level A<sub>0</sub> defined by the metamodeling hierarchy.

**Metamodeling hierarchy.** Fig. 12.b represents the connections among all instances of components used at  $A_0$  level with their component types at  $A_1$  level, and the connections among all component types of the architecture level ( $A_1$ ) with their meta-component defined in C3 metamodel. Those connections are realised using IOC connector.

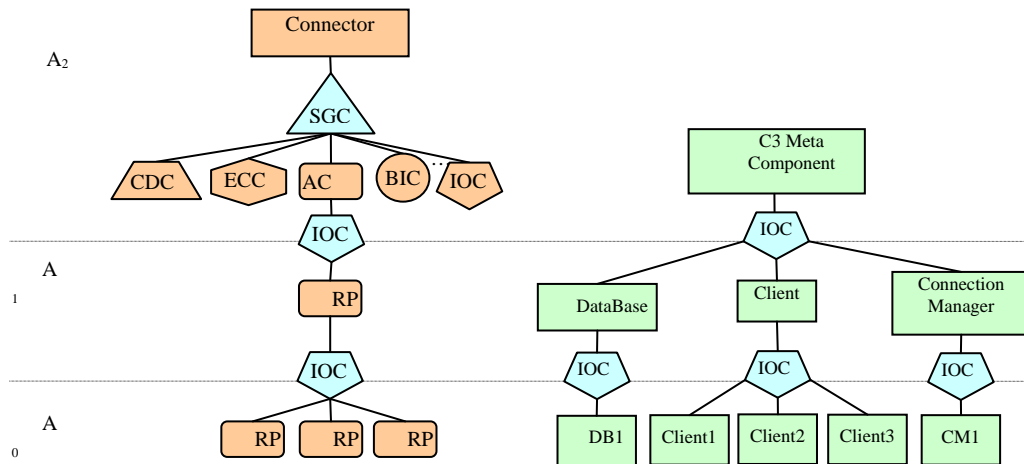


Fig.12. a. Conceptual hierarchy

b. Metamodeling hierarchy

## 5 Conclusion

In this paper we have defined a metamodel called C3 to describe software architecture and to reason about software architecture from different perspective view. The core elements of C3 are component, connector, and configuration. Elements are assembled using their interfaces. Syntactic and semantic corrections are carried out using respectively interfaces-matching and protocols-matching. Perspective views are defined using mainly, SH to describe the structural, BH to describe the behaviour function, CH to describe element libraries, and we use the MH to show how we can operate on C3 metamodel. Each hierarchy is supported by explicit mechanisms to provide the different form of connections required. We have defined six types of connector to deal with different types of connections. SH uses CDC, ECC, and AC. BH uses CDC, BIC, and AC. CH uses the SGC and finally MH uses the IOC.

## References

1. Amirat, A., Oussalah, M., and Khammaci, T.: Towards an Approach for Building Reliable Architectures. In Proc. of IEEE IRI'07, Las Vegas, USA, Pp 467-472, August 2007.
2. Garlan, D., Monroe, R.T., and Wile, D.: Acme: Architectural Description Component-Based Systems, Foundations of Component-Based Systems. Univ. Press, 2000, pp. 47-68.
3. Lanoix, A., Hatebur, D., Heisel, M., and Souquière, J.: Enhancing Dependability of Component-Based Systems. Ada-Europe'07, pp. 41-54, 2007.
4. Medvidovic, N., Dashofy, E., and Taylor, R.N.: Moving Architectural Description from Under the Technology Lamppost. IST Journal pp. 12-31. Vol. 49, No. 1, 2007.
5. OMG: Unified Modeling Language: Infrastructure. [Electronic Version] from <http://www.omg.org/docs/formal/07-02-06.pdf>, 2007.
6. Pinto, M., Fluentes, L., and Troya, M.: A Dynamic Component and Aspect-Oriented Platform. The Computer Journal Vol.48 No. 4, Pp. 401-420, 2005.
7. Smeda, A., Oussalah, M., & Khammaci, T.: MADL: Meta Architecture Description Language. 3rd ICIS International conference on Software Engineering Research, Management & Applications, SERA'05, Michigan, USA, Pp. 152-159, August 2005.
8. Taylor, R. N., Medvidovic, N., Anderson, K. M., Whitehead, JR., E. J., Robbins, J. E., Nies, K. A., Oreizy, P., and Dubrow, D. L.: A component- and message-based architectural style for GUI software. IEEE Trans. Sof. Eng. 22, 6, pp.390-406, June, 1996.

## Une Architecture basée Agent pour la Gestion de l'Interopérabilité des Processus Métiers

Yasmina Rahmoune<sup>1</sup>, Mahmoud Boufaïda<sup>2</sup>

<sup>1</sup> Département d'Informatique  
Centre Universitaire Cheikh Larbi Tébessi de Tébessa

<sup>2</sup> Laboratoire Lire  
Département d'Informatique  
Université Mentouri de Constantine  
rahmoune\_yasmina@yahoo.fr, boufaïda\_mahmoud@yahoo.fr

**Résumé :** L'interopérabilité est un aspect important dans la collaboration des applications hétérogènes, distribuées et complexes. L'évolution du Web présente un milieu favorable à la naissance d'une nouvelle technologie qui permet l'interopérabilité, facilitant les échanges B2B et l'intégration d'applications. Il s'agit des services Web. Dans ce contexte, nous avons proposé une architecture basée agent qui permet d'établir une interopérabilité « technique, opérationnelle et sémantique » entre les processus métiers inter-entreprises. Cette architecture est fondée sur les services Web sémantiques. En utilisant la description des services Web décrite par l'OWL-S, deux protocoles ont été définis; l'un lié à la publication et l'autre lié à l'invocation des services Web par les processus métiers. Ces derniers sont gérés par un moteur workflow. Nous avons exploité un ensemble d'outils pour valider l'architecture proposée.

**Mots clés :** interopérabilité, processus métier, agent, OWL-S, OWL Parser.

### 1 Introduction

Au fil du temps, les environnements informatiques sont devenus plus complexes et plus hétérogènes en raison de la diversité des besoins des clients et des innovations grandissantes de l'industrie informatique. Les infrastructures publiques et privées du secteur informatique sont de plus en plus multiplateformes, multifournisseurs et distribuées à grande échelle.

La cohabitation des applications avec des technologies différentes est devenue une activité critique du point de vue de la garantie de la qualité des services et de l'efficacité organisationnelle et économique des entreprises. Actuellement, l'interopérabilité est un aspect aussi important que la sécurité et la fiabilité pour la gestion des systèmes informatiques ainsi que des environnements de production.

L'interopérabilité a pour objectif de permettre à plusieurs systèmes logiciels de communiquer et de coopérer afin de minimiser les coûts d'intégration de ces systèmes [1].

Aujourd'hui, la majorité des entreprises sont ouvertes sur le Web. L'évolution de ce dernier présente un milieu favorable à la naissance des services Web qui permettent l'interopérabilité et facilitent les échanges B2B et l'intégration d'applications. Le workflow traditionnel d'inter-entreprise est lié à EDI (Electronic Data Interchange) et récemment à ebXML (electronic business XML) [2].

Actuellement, les problèmes posés sont liés à l'amélioration des mécanismes d'interopérabilité et à l'accroissement du taux de changement tout en assurant les axes d'interopérabilité « technique, opérationnelle et sémantique ».

L'utilisation classique des services Web fournit une interopérabilité lourde et syntaxique. Pour résoudre ce problème, plusieurs solutions sont proposées. Nous basons sur l'interopérabilité entre les processus métier<sup>3</sup> qui sont un ensemble d'activités reliées les unes aux autres pour réaliser un objectif bien défini. Dans ce contexte, pour améliorer les techniques d'interopérabilité entre les processus métiers automatisés par un moteur workflow, nous proposons une architecture basée agent permettant de gérer l'interopérabilité des processus métiers en utilisant la description OWL-S<sup>4</sup> pour les services Web. Ces processus peuvent être indépendants, s'exécutant simultanément dans des environnements différents et hétérogènes qui appartiennent à des partenaires différents.

---

<sup>3</sup> Traduction française de la notion de Business Process.

Synonymes: Procédure, Processus d'affaire, Processus opérationnel, Processus d'entreprise.

<sup>4</sup> OWL-S: Ontology Web Language for Services.

Dans la seconde section, nous évoquerons brièvement quelques travaux dans le domaine de l'interopérabilité. Dans la troisième section, nous présenterons un aperçu global du travail réalisé, puis une étude détaillée ainsi que les composants de notre architecture. Dans la dernière section, nous terminerons par une conclusion et des perspectives.

## **2 Travaux avoisinants**

Longtemps étudiée, l'interopérabilité est de plus en plus requise du fait du déploiement d'un grand nombre d'applications sur Internet et de leurs potentiels d'interaction et de coopération.

Aujourd'hui, les besoins des clients sont variés et ont augmenté avec l'évolution des technologies. Les processus métiers d'une entreprise deviennent de plus en plus complexes. Ces processus peuvent être très longs et lourds et peuvent dépasser la frontière de l'entreprise. Ils ont besoin d'une collaboration avec d'autres processus indépendants s'exécutant simultanément dans des environnements différents, hétérogènes, qui appartiennent à des partenaires différents. Des mécanismes sont nécessaires pour interopérer entre eux d'une manière efficace. Ils peuvent être très dynamiques afin de mieux répondre aux besoins spécifiques des clients et aux changements des relations avec des partenaires. Pour améliorer ces mécanismes d'interopérabilité, plusieurs travaux ont été réalisés, parmi lesquels nous citons :

### **❖ L'interopérabilité des systèmes multi-agents à l'aide des services Web**

T.Melitti et al proposent un cadre conceptuel et architectural pour l'interopérabilité des systèmes multi-agents (SMA) hétérogènes [1]. Dans leur modèle, les agents publient leurs capacités en tant que services Web pouvant être utilisés par d'autres agents et ce indépendamment des caractéristiques conceptuelles (architecture, modèle d'interaction, etc.) et techniques (plate-forme, langage de programmation, etc.) de ces agents.

### **❖ L'intégration des agents dans les workflows**

L'approche proposée par Buhler et Vidal [3] consiste à utiliser une description de processus en BPEL4WS<sup>5</sup> pour imposer un ordre social initial à une collection d'agents. Comme BPEL4WS décrit les rapports entre les services Web dans le workflow, Les agents représentant des services Web sauraient leurs rapports a priori [3] [4].

L'établissement de workflow par un système multi-agents peut être vu comme acte de solution des problèmes de coopération [5] [6].

### **❖ Utilisation des agents VDL<sup>6</sup> pour l'interopérabilité**

Yasmine Charif et Nicolas Sabouret proposent un modèle de programmation d'agents assistants capables d'interopérer avec des services Web et d'autres agents assistants en s'appuyant sur un modèle dialogique de requêtes formelles permettant de raisonner sur leur propre code en cours d'exécution [7].

Dans leurs travaux de recherche, ils définissent des services Web sémantiques, actifs (ASWS), c'est-à-dire des agents assistants (en ligne dans l'Internet) capables d'interagir à la fois avec des utilisateurs humains ainsi qu'avec des services Web. Du point de vue de l'utilisateur, les ASWS sont des pages Web actives. Mais contrairement aux services classiques qui sont uniquement réactifs, les ASWS fonctionnent à la fois en mode réactif et proactif.

Nous pouvons résumer ces travaux par la nécessité d'ajouter les agents entre les processus métiers et les services Web pour réaliser cette interopérabilité. Pour aboutir à une interopérabilité flexible et adaptative entre les processus métiers et qui vérifie les trois axes d'interopérabilité, nous proposons une architecture basée agent pour gérer l'interopérabilité de ces processus.

---

<sup>5</sup> BPEL4WS: Business Process Execution Language for Web Services.

<sup>6</sup> VDL: View Design Language.

### 3 L'architecture proposée

Nous présentons dans cette section une démarche complète pour réaliser l'interopérabilité entre les processus métiers.

#### 3.1 Aperçu global

De nos jours, les processus peuvent être très longs et complexes et peuvent même dépasser la capacité de l'entreprise. Pour répondre aux exigences des clients, ils font collaborer avec d'autres processus.

Notre travail est fondé sur l'interopérabilité des activités entre les processus métiers indépendamment des caractéristiques conceptuelles et techniques de ces processus intra et inter-entreprises.

Nous utilisons les mécanismes du service Web, en considérant les activités invoquées ou publiées des services Web. L'interopérabilité entre les processus métiers par l'utilisation des services Web directement est lourde et pauvre sémantiquement. Pour résoudre ce problème nous proposons d'utiliser des agents pour accéder aux services par les processus et la description sémantique OWL-S pour les services Web.

Chaque entreprise a un UDDI<sup>7</sup> local, indexant notamment les services Web mis à la disposition par cette entreprise (ou ses processus métiers). Cet UDDI local est divisé en deux catégories. La première contient les services privés, invocables seulement à partir des processus internes (c'est-à-dire de la même entreprise). La seconde contient les services publics, invocables par tous les processus internes ou externes.

Nous considérons que chaque entreprise a un agent nommé « Agent-Entreprise » qui représente l'entreprise à l'extérieur et qui prend la responsabilité d'établir la communication avec les autres entreprises, de publication et d'invocation des services Web de cette entreprise (les activités ou les sous processus).

Pendant la découverte des services, nous proposons un agent du service Web nommé « Agent-SW » qui reçoit les requêtes envoyées par les agents des entreprises et fait la recherche du service. Pour trouver le service adéquat, il prend également en considération l'aspect sémantique. Nous présenterons le fonctionnement de ces agents dans les sections 3.5 et 3.6.

#### 3.2 Architecture interne de l'entreprise

Chaque entreprise est créée pour accomplir un certain nombre d'objectifs. Pour les réaliser, elle déclenche un ensemble de processus métiers. L'architecture interne proposée d'une entreprise est définie dans la figure 1.

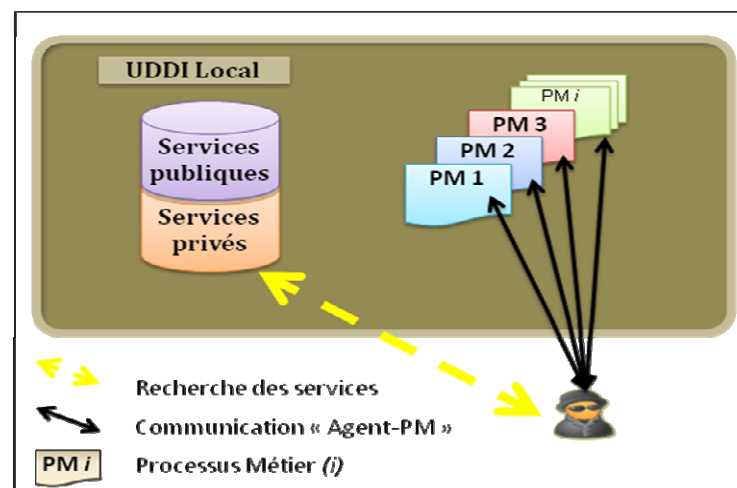


Fig. 1. Architecture interne d'une entreprise.

<sup>7</sup> UDDI: Universal Description, Discovery and Integration.

Un processus métier est composé d'un ensemble d'activités qui sont prises en charge par des participants vérifiant certaines règles. Les participants peuvent être des humains, des machines ou d'autres applications. Nous avons axé notre travail sur l'interopérabilité des processus métiers intégrant les services Web, sachant que ces services Web sont publiés par d'autres entreprises (processus métiers), c'est-à-dire qu'un ou plusieurs services Web interviennent dans l'exécution du processus métier en tant qu'applications invoquées. L'invocation est établie par les deux agents proposés : l'agent d'entreprise et l'agent du service Web.

### 3.2.1 Modes de communication

Dans notre architecture, nous avons considéré qu'il existe deux niveaux de communication :

- Intra-entreprise: l'échange des activités est établi entre les processus métiers de la même entreprise par l'utilisation de l'UDDI local.
- Inter-entreprises: l'échange des activités est établi entre les processus métiers d'entreprises différentes.

### 3.3 Interaction des agents avec les services Web

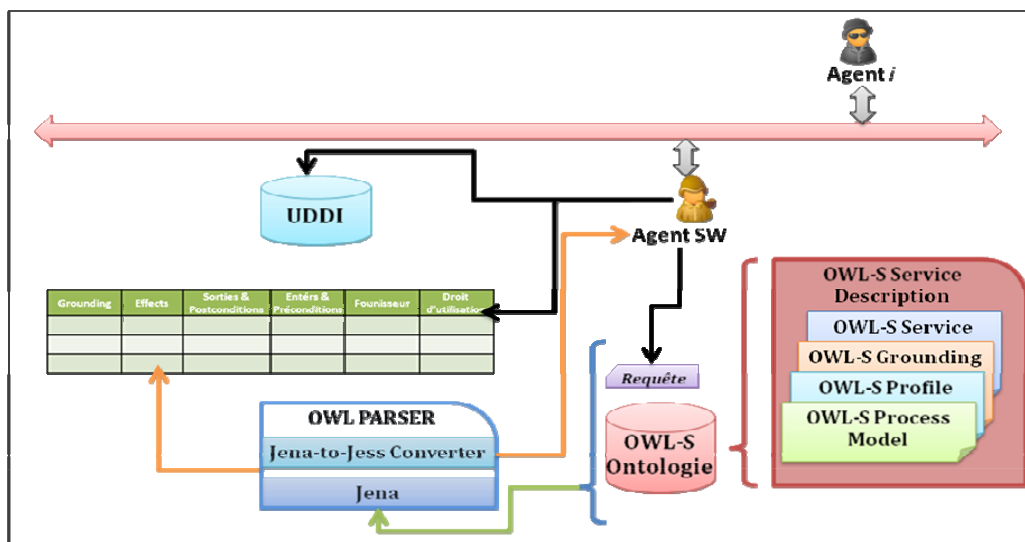


Fig. 2. Processus d'interaction de l'agent service Web avec les ontologies et l'UDDI.

Ce schéma présente l'interaction entre les agents et les services Web, tel que :

- L'agent « Agent-SW » est prêt pour recevoir les requêtes à partir des autres agents (les agents des entreprises). Puis il va chercher les services demandés.
- Ontologie de description des services Web « OWL-S ontologies ».
- L'analyseur OWL « OWL Parser », le résultat de cet analyseur est une base de données contenant les services qui peuvent répondre à notre requête.

### 3.4 Fonctionnement de l'agent du service Web

L'agent du service Web reçoit la requête d'invocation d'un service envoyée par l'agent d'entreprise, puis il établit le lien entre les processus des entreprises et les ontologies des services Web pour invoquer des services. Cet agent doit avoir les caractéristiques suivantes:

1- Il doit connaître l'ontologie liée au domaine de l'application. A partir de la requête reçue l'agent extrait le domaine de l'ontologie qu'il va utiliser.

Dans notre travail, nous nous sommes inspiré l'analyseur développé dans [8]. Le DAML-S et le DAML Parser ont été modifiés par leurs successeurs l'OWL-S et le OWL Parser parce que l'OWL est basé sur le DAML plus d'autres fonctionnalités.

2- Il peut analyser et interpréter des caractéristiques d'OWL-S pour rechercher les services employés par l'application. Ce rôle est assuré par l'analyseur OWL « OWL Parser ». Ce dernier est basé sur l'analyseur de Jena « Jena Parser » [9], qui lit des spécifications de OWL-S des services Web, puis les



traduit dans un ensemble de prédicats, et les passe au Jena-to-Jess, ce dernier va les moduler. Il est basé sur l'épreuve de théorème de Jess [10], ce qui est employé pour mettre en application une sémantique d'OWL-S. Le résultat de cette étape est une base de données qui contient les services existants suivant le domaine et le contenu des ontologies (voir la figure 2). Pour obtenir une recherche autonome et efficace, l'Agent-SW maintient cette base de données contenant les différentes propriétés du service lié à leur découverte et à leur choix. Ces propriétés sont extraites à partir du profil de service « service profile » d'OWL-S et du modèle de service « service model » d'OWL-S. En effet le service grounding d'OWL-S est insuffisant pour l'invocation de service Web. Pour cela, nous proposons l'utilisation du WSDL, qui est le meilleur dans ce niveau [11].

### **3.5 Protocoles utilisés**

Les processus d'invocation et de publication sont présentés dans les deux protocoles suivants. Dans cet article nous avons mis l'accent sur le protocole d'invocation.

#### **3.5.1 Protocole d'invocation**

- Lorsque le responsable d'un processus métier veut chercher et invoquer des services, il envoie une requête à l'agent d'entreprise "AgentEi".
- L'agent "AgentEi" interroge en premier lieu l'UDDI local. Si ses besoins sont vérifiés, on terminera le protocole d'invocation par l'appel de ces services, sinon l'agent de l'entreprise envoie sa requête à l'agent de service Web "Agent-SW".
- L'Agent-SW reçoit la requête, il l'analyse et extrait ses propriétés. Il envoie la requête et l'ontologie de description des services à l'OWL Parser. Ce dernier va construire une base de données (à partir des étapes citées précédemment) qui contiendra les services pouvant répondre à la requête. Puis l'Agent-SW va commencer la recherche du service dans cette base de données (voir la figure 2) suivant un ensemble des critères de recherche. Nous supposons que les critères de recherche sont les entrées et les préconditions d'une part, les sorties et les postconditions du service d'autre part.
- Après la recherche, l'Agent-SW découvre deux cas :
  - Service trouvé: dans ce cas, la liste des adresses de ces services sera directement envoyée à l'AgentEi.
  - Service non trouvé: aucun service trouvé dans la base de données pour l'AgentEi. Dans ce cas, l'Agent-SW renvoie la requête à l'UDDI (recherche classique par un agent). Puis il va sélectionner la liste des services qui peuvent satisfaire leurs besoins. A la fin, il envoie la liste trouvée à l'AgentEi.
- L'agent AgentEi va communiquer directement avec les agents des entreprises spécifiées « dans la liste envoyée par l'Agent-SW ».
- L'AgentEi négocie avec les agents jusqu'à la satisfaction de leur besoin, par l'utilisation d'un protocole de négociation. Le protocole de négociation que nous suggérons pour sélectionner le service adéquat est FIPA-CNP (Contract Net Protocol).
- Vers la fin de la négociation, l'agent choisi déclenche l'exécution du service sélectionné.
- Après l'exécution de service, l'agent demandeur "AgentEi" va signaler à l'agent d'entreprise la satisfaction de ses besoins, ensuite il fera la mise à jour de ses connaissances.
- Si l'AgentEi met son accord sur un des services d'UDDI, il envoie l'adresse de ce service à l'Agent-SW. Ce dernier enregistre ce service et met sa nouvelle description sémantique dans l'ontologie des services Web, afin de l'employer dans les cas similaires. Cette démarche rend notre système plus adaptatif et plus autonome avec la progression du temps.

#### **3.5.2 Protocole de publication**

- Lorsque le responsable du processus métier décide de publier une activité, il envoie sa description à l'agent d'entreprise AgentEi tout en précisant le type de cette activité «service», privée ou publique.
- A partir de la description de l'activité « service », l'agent d'entreprise "AgentEi" va la publier dans l'UDDI local selon son type.
- Si le type de service est public, la description du service est publiée à travers l'agent service Web dans l'ontologie OWL-S.

#### **4. Conclusion et perspectives**

Afin de maximiser l'intégration des systèmes hétérogènes et de favoriser une interopérabilité à moindre coût, il est nécessaire de minimiser les exigences en terme de spécifications.

Ce travail a permis de réaliser une interopérabilité automatique entre les processus métiers inter-entreprises par l'utilisation des agents. Il vérifie aussi tous types d'interopérabilité l'utilisation des ontologies des descriptions des services Web « OWL-S ». Nous avons élargi le champ de l'interopérabilité entre les processus métiers quelque soit leurs participants.

Dans un futur proche, nous estimons pouvoir former un Web interactif et intelligent grâce à l'intégration les technologies des services Web et l'utilisation de l'aspect sémantique dans la recherche. A cet effet, il est nécessaire de créer des ontologies standard classifiées par domaines qui seront utilisées durant la description sémantique des services afin d'éliminer complètement le problème d'ambiguïté du sens.

#### **Références**

1. Melliti, T.: Interopérabilité des services Web complexes. Application aux systèmes multi-agents. Thèse de Doctorat de l'Université Paris IX Dauphine. (décembre 2004)
2. Shuying, W., Weiming, S. et Qi H.,: An agent-based Web service workflow model for inter-enterprise collaboration. Integrated Manufacturing Technologies Institute, National Research Council of Canada, 800 Collip Circle, London, Ont., Canada N6G 4X8. (2006)
3. Buhler, P.A, Vidal, J.M. et Verhagen, H.: Adaptive Work-flow = Web Services + Agents. Proc. Int'l Conf. Web Services, CSREA Press, pp. 131-137. (2003)
4. Vidal, J.M., Buhler, P.A. et Christian, S.: Multiagent Systems with Workflows. JANUARY-FEBRUARY 2004 Published by the IEEE Computer Society 1089-7801/04/ (2004)
5. Buhler, P.A.: A Software Architecture for distributed Workflow enactment with Agents and Services Web. Thèse de Doctorat Philosophique Department of Computer Science and Engineering College of Engineering and Information Technology University of South Carolina. (2004)
6. Buhler, P.A. et Vidal, J.M.: Towards Adaptive Workflow Enactment Using Multiagent Systems. L, Information Technology and Management journal. (2005)
7. Charif, Y. et Sabouret, N.: Programmer des agents assistants interopérables dans le Web sémantique. (octobre 2004)
8. Brahimi, M., Boufaïda, M. et Seinturier, L.: Integrating Web Services within Cooperative Multi Agent Architecture. Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW 2006) 0-7695-2522-9/06 \$20.00 © 2006 IEEE.(2006)
9. B. McBride: JenaSemantic Web Toolkit. <http://www.hpl.hp.com/semweb/jena.htm>
10. E. Friedman-Hill,: Jess: The rule engine for the Java Platform. <http://www.herzberg.ca.sandia.gov/jess/>
11. Christensen, E., Curberam, F., Meredith, G. et Weerawarana, S.: Web Service Description Language (WSDL). <http://www.w3.org/TR/wsdl>, (2001).

## Approche distribuée pour la classification supervisée

Moez Ben Haj Hmida<sup>1</sup>, Yahya Slimani<sup>1</sup>

<sup>1</sup> Faculté des Sciences de Tunis, Campus Universitaire, 1060 Tunis, Tunisie  
{moez.benhajhmida, yahya.slimani}@fst.rnu.tn

**Résumé.** La classification est une technique de fouille de donnée largement utilisée dans des domaines critiques comme l'analyse de risques financiers, la biologie, la gestion de réseaux de communication, etc. Dans ce papier, nous proposons une approche de classification supervisée pour de larges ensembles de données distribués. Cette approche se base sur une migration de code (algorithmes de classification) plutôt que sur la migration de données. A travers un certain nombre d'expériences, nous montrons que notre approche permet d'obtenir de meilleures performances par rapport à une approche centralisée et une approche parallèle.

**Mots clés:** Classification supervisée, Bases de données distribuées, Algorithmes parallèles, Réseaux de neurones.

### 1 Introduction

La classification est une technique de fouille de donnée largement utilisée dans des domaines très diversifiés pour extraire de la connaissance cachée dans de larges ensembles de données. Les algorithmes de classification actuels reposent sur un ensemble de théories, modèles, techniques et outils, parmi lesquels nous pouvons citer les statistiques, l'induction, les réseaux de neurones [4] ou les algorithmes génétiques [5]. Malgré cette diversité et pluralité de méthodes et d'outils, les algorithmes actuels souffrent de deux éléments : (i) la qualité de la classification obtenue qui laisse encore à désirer; et (ii) leurs performances.

Pour améliorer les performances des classifieurs, il est possible d'utiliser des techniques de parallélisation [3,8]. L'exploitation du parallélisme dans les algorithmes de classification suit deux stratégies principales : le parallélisme de données et le parallélisme de tâche. Dans le parallélisme de données, la base de données est divisée sur tous les processeurs et chaque processeur exécute localement la même tâche sur ses propres données. Dans le parallélisme de tâche, différentes tâches sont exécutées concurremment sur différentes partitions mais les processeurs doivent accéder à la base de données entière. Dans ce cas-ci, la base de données est explicitement répliquée (mémoire distribuée) ou partagée par les différents processeurs (mémoire partagée). Le parallélisme hybride est la combinaison des deux stratégies [2].

Cependant si la taille de cette base est très importante ou si elle est largement distribuée, sa duplication peut engendrer un surcoût de communication qui peut dégrader les performances d'un système de classification. Il est donc préférable dans ce cas d'utiliser le parallélisme de données.

Ainsi, la combinaison de techniques de parallélisation et de distribution pourra contribuer à améliorer, de manière sensible, les performances des classifieurs. Dans ce papier, nous proposons une approche de classification supervisée basée sur la distribution de données.

Le reste du papier est organisé comme suit. La section 2 fait un rappel sur les techniques parallèles de classification supervisée. L'approche proposée sera décrite dans la section 3. La section 4 présente et analyse quelques résultats d'expérimentations qui permettent de valider l'approche proposée. Enfin, la section 5 conclut ce papier et liste quelques perspectives de prolongement de nos travaux.

### 2 Classification supervisée

La classification supervisée a pour tâche l'identification de la classe d'appartenance d'un objet à partir de certains traits (caractéristiques) descriptifs. Ce type de classification nécessite deux phases: l'apprentissage et la classification. La première phase consiste à créer un modèle de classification spécifique à partir d'un jeu de données d'apprentissage. Ensuite, la qualité du modèle de classification est

évaluée sur des données de test. La phase de classification (ou prédiction) est l'application du modèle, trouvé lors de la première phase, sur de nouvelles données.

## 2.1 Techniques parallèles de classification supervisée

Pour améliorer les performances des classifieurs, il est possible d'utiliser des techniques de parallélisation. La majorité des algorithmes de classification parallèles se basent sur le parallélisme de données distribuées. Dans ce type d'algorithmes, le même classifieur est appliqué à chaque partition de la base de données afin de générer un modèle local. Par la suite, tous ces modèles seront agrégés par une stratégie de combinaison pour produire un modèle final [1].

Un framework de meta-learning a été proposé [6] pour répondre au problème de classification des bases de données distribuées. En premier lieu, pour chaque site un classifieur de base est construit à partir de la base d'entraînement locale. Les classifieurs ainsi construits sont rassemblés dans un site central où leurs prédictions sont évaluées sur une base de validation séparée pour former des données meta-niveaux. En final, le classifieur global (meta-classifieur) est entraîné sur les données meta-niveaux. Le rôle du meta-classifieur est la combinaison ou la fusion des différentes prédictions qui forment les données meta-niveaux. L'exécution d'une méthode meta-learning peut se résumer dans les étapes suivantes: (i) générer un classifieur de base sur chaque site en utilisant un algorithme de classification, (ii) regrouper les classifieurs de base dans un site central. Construire des données de méta-niveau en générant les prédictions de ces classifieurs sur une base de données de validation, et (iii) générer un méta-classifieur (classifieur final) à partir de la méta-base, formée des prédictions des différents classifieurs ainsi que la base de validation.

La figure 1 illustre les différentes étapes de meta-learning appliquées à une base de données distribuées sur deux sites.

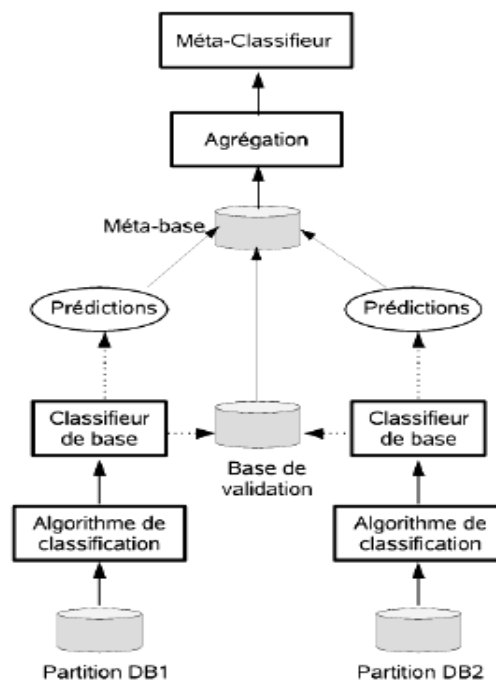


Fig. 1. Meta-learning à partir d'une base de données distribuée.

## 3 Modèle proposé

Dans cette section, nous proposons une approche de classification supervisée qui s'inspire de l'algorithme génétique parallèle appelé modèle d'îlots [5]. Ce modèle crée un ensemble de solutions de départ qu'il divise en îlots. Chaque solution est considérée comme un individu qui migre d'un îlot à un autre. Le but de la migration est l'amélioration des individus en subissant des croisements et des mutations.

Par analogie au modèle d'îlots, et dans le cas d'une base de données géographiquement distribuée, chaque fragment de la base sera considéré comme un îlot. Un individu (solution) sera un modèle de classification supervisée. Initialement, un modèle est créé sur chaque fragment. En migrant d'un site à un autre, le modèle doit s'adapter aux données de l'îlot d'accueil. Les modèles sont construits en utilisant l'algorithme des réseaux de neurones à rétro-propagation du gradient [4].

Dans ce qui suit nous allons décrire le problème à traiter et nous détaillerons ensuite le modèle que nous proposons.

### 3.1 Position du problème

L'avancée technologique des réseaux informatiques et de l'Internet ont permis la génération (automatique) de larges volumes de données distribuées, qui renferment une source de connaissances cachée, très riche et très utile pour les décideurs. L'extraction de cette connaissance nécessite la mise en oeuvre de techniques de fouille de données ou de datamining. Dans le cas d'un système de datamining distribué, alors la première phase d'un processus d'extraction consiste à définir le mode de partitionnement et de distribution des données.

Dans ce papier, nous nous intéressons au cas d'une base de données distribuée homogène. Le traitement que nous proposons, pour ce type de base, se fera d'une manière distribuée sans déplacement (migration) de données.

### 3.2 Méthodologie

Considérons une base de données distribuée  $D$  partitionnée sur  $N$  sites  $S_i$ ,  $i \in \{1, \dots, N\}$ , répartis géographiquement. Au lieu de déplacer ces partitions sur un seul site pour les traiter d'une manière centralisée, nous proposons de construire un modèle local qui parcourt tous les sites pour aboutir à un modèle global.

La figure 2 montre les différentes étapes de l'approche proposée. C'est en effet une répétition de deux étapes essentielles qui sont l'apprentissage et la migration. L'étape d'apprentissage est l'application d'un algorithme de classification à la base locale de chaque site afin de produire un modèle local par site. L'étape de migration consiste à construire un anneau entre les différents sites et à faire migrer les modèles locaux aux sites voisins. A chaque nouvelle étape d'apprentissage, chaque site exécute le même algorithme de classification pour entraîner le modèle reçu aux données la base locale. A chaque nouvelle étape de migration, l'anneau sera créé dans un nouvel ordre respectant un schéma de migration prédéfini.

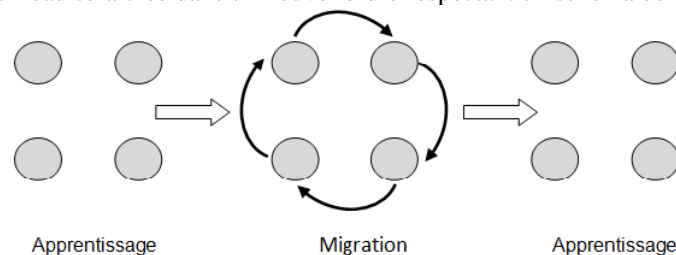


Fig. 2. Modèle d'apprentissage par migration de code.

Généralement, un classifieur est construit à partir d'un modèle initial nul (arbre de décision) ou d'un modèle construit aléatoirement (réseaux de neurones). Ceci signifie que l'étape d'apprentissage se fait sans connaissances préalables. Par contre, dans notre approche, chaque modèle est construit à partir du modèle précédent. Nous appelons cette forme d'apprentissage, le *ré-apprentissage*.

Les réseaux de neurones sont parfaitement adaptés à ce type de traitement grâce à l'aspect itératif des algorithmes qu'ils utilisent. Malheureusement ces réseaux sont connus par leur sensibilité aux états initiaux. L'architecture du réseau, nombre de couches internes et nombres de noeuds, peut avoir des effets sur les performances du modèle construit. Ils sont également sensibles à l'ordre de traitement des instances. Pour remédier à ces problèmes, nous proposons d'utiliser une approche basée sur le parallélisme et la distribution. Ainsi, sur chaque site le même algorithme est appliqué à la partition locale. Les modèles ainsi construits vont migrer parallèlement d'un site à un autre pour produire  $N$  modèles globaux. Chaque modèle est caractérisé par son architecture et un ordre de parcours des partitions. Le choix du meilleur modèle sera basé sur la qualité des performances obtenues pour chaque modèle.

### 3.3 Algorithme

Notre algorithme prend en entrée une base de données distribuée géographiquement sur  $N$  noeuds contenant chacun une partition  $DBi$  et une base de test  $BT$  construite à partir des différentes partitions. Initialement, un réseau de neurones de poids aléatoires est créé sur chaque noeud. Nous avons ainsi un modèle initial sur les différents noeuds. Chaque modèle sera ensuite entraîné par l'algorithme à rétropropagation du gradient [4] sur la partition locale puis évalué sur la base de test. Selon le schéma de migration, chaque modèle est envoyé vers le noeud suivant où il va cumuler l'apprentissage par entraînement sur la partition du noeud récemment visité. Le processus est réitéré par chaque modèle en migrant d'un noeud à un autre jusqu'à atteindre le noeud de départ. Le pseudo-code de l'algorithme est le suivant :

---

**Algorithme 1.**


---

**Pour**  $i$  de 1 à  $n$  (chaque noeud)

**Entrée :**

$DBi, BT$

$Si$  : schéma de migration

**Sortie :**

Modèle  $M$

$M$  := Réseau de neurones de poids aléatoires

$S$  :=  $Si$

**Pour**  $k$  de 1 à  $n$

Entraîner  $M$  sur  $DBi$

Tester  $M$  sur  $BT$

Envoyer  $M$  et  $S$  à  $S[k+1]$

Envoyer  $S$  à  $S[k+1]$

Recevoir  $M$  et  $S$

Elire le meilleur modèle

**Fin**

---

Cet algorithme sera exécuté en parallèle par chaque noeud suivant un schéma de migration préalablement défini. Ce schéma est en fait un tableau circulaire contenant les labels des noeuds afin de décrire l'ordre de parcours des noeuds. Il faut préciser ici que chaque noeud aura un schéma de migration distinct et qu'il existe ( $N!$ ) schémas possibles. L'exécution de tous les schémas de migration possibles sera très coûteuse en temps de calcul et nécessitera plusieurs exécutions de l'algorithme. Dans notre approche, seuls  $N$  ordres de parcours sont pris en compte. Ainsi l'algorithme est exécuté une seule fois en parcourant une partie, choisie aléatoirement, de l'ensemble des solutions possibles.

La Figure 3 décrit les différentes étapes de l'algorithme proposé sur une base composée de quatre partitions. Un modèle ( $M_i$ ) est construit par un algorithme d'apprentissage supervisé à partir de la première partition ( $P_i$ ). Le modèle ( $M_i$ ) est ensuite réutilisé dans la phase d'apprentissage à partir de la partition ( $P_j$ ).

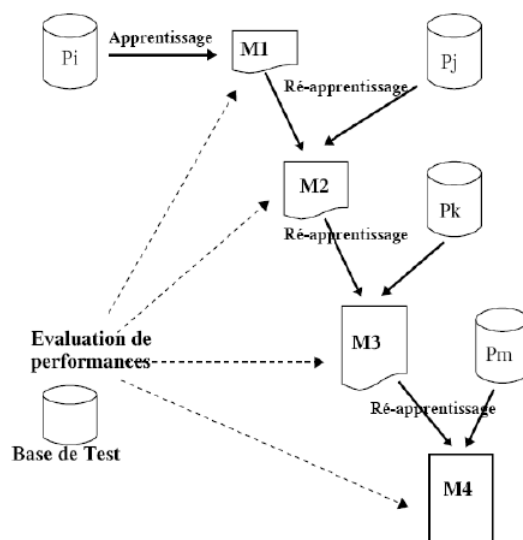


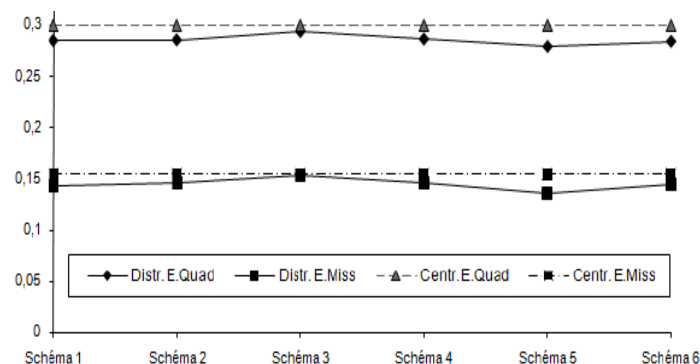
Fig. 3. Apprentissage à partir d'une base de données distribuée.

Maintenant, pour comparer les performances de notre algorithme nous allons appliquer ce modèle à un algorithme d'apprentissage basé sur les réseaux de neurones.

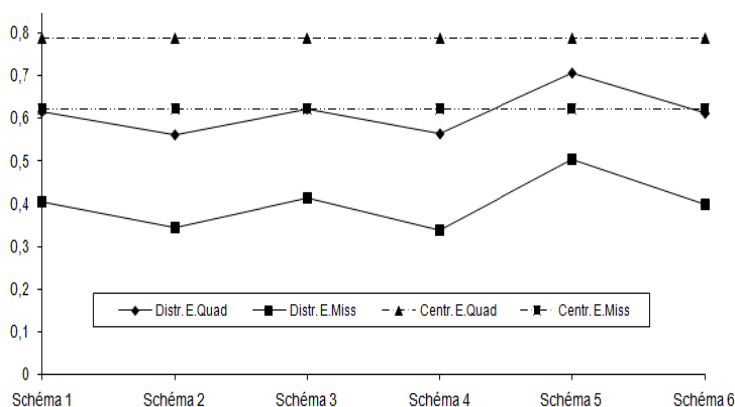
#### 4 Evaluation expérimentale

Dans cette section, nous discutons de quelques résultats d'expérimentations que nous avons effectuées pour comparer les performances de notre approche par rapport à une approche centralisée. Ces expérimentations ont été réalisées sur un PC Duo-Core de fréquence 1,6 GHz avec 1 GB de RAM fonctionnant sous l'environnement Fedora4 et utilisant la machine virtuelle Java 1.5. L'algorithme d'apprentissage utilisé est l'algorithme de rétro-propagation du gradient de la plateforme Weka 3.5.3 [7] que nous avons étendu pour implémenter notre approche. Pour comparer notre approche avec une approche centralisée, nous avons choisi les bases Waveform et Colic du site UCI (<http://kdd.ics.uci.edu/>). La base Waveform est composée de 5000 instances, 3 classes et 40 attributs. Elle sera divisée en une base de test qui représente 10% de la base et le reste comme base d'entraînement. La base Colic contient 27 attributs, 2 classes et 368 instances dont 300 d'entraînement et 68 de tests. Pour expérimenter notre approche, chaque base d'entraînement sera divisée en 4 partitions égales. Les mesures de performances à considérer sont l'erreur quadratique et l'erreur de classification (misclassification). L'erreur quadratique est le carré de la différence entre valeur prédite et valeur attendue alors que l'erreur de classification est la moyenne pondérée des erreurs sur les descriptions.

Les réseaux de neurones utilisés ont une seule couche cachée. Le nombre de nœuds est égal à la somme du nombre d'attributs et du nombre de classes, ou la moitié de cette somme. Le schéma de migration des modèles. Notre base étant partitionnée en 4 fragments a, b, c et d, nous avons considéré 6 schémas possibles: (a b c d), (b c d a), (c d a b), (d a b c), (a d b c) et (a c d b). Pour chaque schéma, un modèle est créé sur le premier fragment puis il migre vers les autres partitions suivant l'ordre ainsi défini. La Figure 4 compare les valeurs de l'erreur quadratique et l'erreur de classification obtenues par notre approche avec celles de l'approche centralisée, par rapport à la base Waveform suivant les 6 schémas définis précédemment. Nous constatons que les performances de notre approche varient selon le schéma de migration mais restent meilleures que l'approche centralisée. Pour la Figure 5, nous constatons que les performances de notre approche surpassent largement celles de l'approche centralisée pour la base Colic qui est assez volumineuse. Cette amélioration de performances est due à la réadaptation du classifieur lors de son parcours des différents fragments de la base, ce qui permet un apprentissage plus général et évite le sur-apprentissage.



**Fig. 4.** Influence de l'ordre de parcourt des partitions sur la précision de classification de notre algorithme pour la base Waveform.



**Fig. 5.** Influence de l'ordre de parcourt des partitions sur la précision de classification de notre algorithme pour la base Colic.

Les figures 4 et 5 montrent bien la sensibilité de la précision de classification entre les différents schémas de parcourt des partitions. En effet, chaque fois qu'un modèle est appliqué à une nouvelle partition ses connaissances sont adaptées aux informations qu'elle contient. Cette adaptation peut améliorer ou dégrader la qualité du classifieur selon l'ordre de parcourt et non pas la partition analysée en début ou à la fin du parcourt.

## 5 Conclusion

Dans ce papier, nous avons proposé une approche distribuée de classification supervisée dans laquelle nous utilisons un modèle de migration du calcul au lieu de la migration des données, qui sont beaucoup plus volumineuses. Cette approche, appliquée à des bases de données distribuées, a permis de faire un apprentissage qui permet d'éviter le surcoût lié au déplacement des données. L'efficacité de notre approche a été comparée par rapport à celle d'une approche centralisée, où la base est stockée sur un seul site pour construire une base de connaissances. Les premières expériences que nous avons menées ont montré, de manière assez significative, le gain en performances obtenu par l'approche proposée dans ce papier. Ces résultats encourageants laissent entrevoir de nouvelles orientations pour améliorer notre approche. Il est envisagé aussi de multiplier les expérimentations notamment pour des bases plus volumineuses. Ainsi, nous pensons qu'il serait intéressant d'utiliser les algorithmes génétiques pour retrouver un réglage optimal des paramètres initiaux. Ce réglage consiste à trouver un nombre de couches internes ainsi que le nombre de noeuds par couches qui améliorent les performances du réseau de neurones. Nous pensons aussi que le schéma de migration peut être fixé par l'algorithme génétique.



## **Références**

1. Chan, P. K., Stolfo, S. J.: On the Accuracy of Meta-learning for Scalable Data Mining, *Journal of Intelligent Information Systems*, Springer, (1997) 5–28
2. Freitas, A.A.: A survey of parallel data mining. 2nd Int. Conference on the Practical Applications of Knowledge Discovery and Data Mining (1998), 287-300
3. Kargupta, H., Chan, P. (Ed): *Advances in Distributed and Parallel Knowledge Discovery*. MIT Press (2000)
4. Michie, D., Spiegelhalter, D., Taylor, C.: *Machine Learning, Neural and Statistical Classification*. Artificial Intelligence. Ellis Horwood (1994)
5. Skolicki, Z., De Jong, K.: Improving evolutionary algorithms with multirepresentation island models. 8th International Conference on Parallel Problem Solving from Nature PPSN VIII (2004)
6. Stolfo, S.: Jam: Java agents for meta-learning over distributed databases. Third International Conference on Knowledge Discovery and Data Mining, (1997) 74–81
7. Witten, I., Frank, E.: *Data Mining: Practical machine learning tools and techniques*. 2<sup>nd</sup> Edition, Morgan Kaufmann, San Francisco (2005)
8. Zaki, M. J., Ho, C. T. (ed.): *Large-Scale Parallel Data Mining*, *Lecture Notes in Computer Science*, Springer, (2000)

## **APPROCHE BOND-GRAPH POUR LA MODELISATION DES SYSTEMES D'INFORMATION (BGSI): NOUVELLE APPROCHE ORIENTEE COMMUNICATION**

ABED BAHTOU B., RAHMOUNI M.K.

Département d'Informatique  
Université d'Oran Es-Sénia  
B.P 1524, Oran El-M'naour Oran (Algérie)

**Résumé:** L'outil Bond-graph (ou graphe de liaison), défini par [PAYTER 1961], formalisé par [KARNOPP, ROSENBERGER, 1975,1983], [BREEDVELD, 1984], se situe comme intermédiaire entre le système physique et les modèles mathématiques qui lui sont associés. Il est très utilisé dans les langages de simulations orientés-objets des systèmes physiques comme un modèle intermédiaire qui sert à générer les équations mathématiques qui décrivent le comportement des systèmes physiques. L'objectif principal de ce travail consiste à faire adapter cet outil à la modélisation des systèmes d'information. Vu que le sujet traité est assez compliqué et touche trois domaines tout à fait différents alors on va essayer de répondre essentiellement à trois questions :

- Traduire les concepts du langage Bond-Graphs en leurs correspondants dans le langage UML
- pour la modélisation des S.I;
- Définir Leurs structures;
- Montrer le point fort de la méthode (Communication).

**Mots-clés :** Bond-Graphs, Systèmes, UML, Structures, Modélisation, Simulation, Langages, orientés objets, Communication.

### **1. Introduction :**

L'outil Bond-graph (ou graphe de liaison), défini par [PAYTER 1961], formalisé par [KARNOPP, ROSENBERGER, 1975,1983], [BREEDVELD, 1984], se situe comme intermédiaire entre le système physique et les modèles mathématiques qui lui sont associés. Il représente à l'heure actuelle la méthodologie de modélisation la plus élaborée en ce qui concerne les systèmes physiques et dynamiques [1, 2, 3,4]. Il est très utilisé dans les langages de simulations orientés-objets des systèmes physiques comme un modèle intermédiaire qui sert à générer les équations mathématiques qui décrivent le comportement des systèmes physiques. L'objectif principal de ce travail consiste à faire adapter cet outil à la modélisation des systèmes d'information. Vu que le sujet traité est assez compliqué et qui touche trois domaines tout à fait différents alors on va essayer de répondre essentiellement à trois questions [1] :

- Traduire les concepts du langage Bond-Graphs en leurs correspondants dans le langage UML pour la modélisation des S.I;
- Définir Leurs structures;
- Montrer le point fort de la méthode (Communication).

D'une autre manière, c'est montrer la faisabilité de l'adaptation.

Ce papier comporte quatre sections qui sont organisées comme suit :

Dans la première section, on va présenter les différents concepts et définitions de la méthode dans les systèmes physiques. Dans la section 2, on donne notre point de vue concernant l'adaptation de la méthode dans les systèmes d'information. La section 3 a pour but de situer cette méthode vis-à-vis d'autres méthodes de modélisation des systèmes d'information (définition des classes et comparaison avec les modèles UMLs). Enfin, je termine ce travail avec une conclusion ou on va citer les perspectives futures.

## 2. Bond-Graphs et systèmes physiques [2, 3, 4]:

Le langage bond-graph est un langage graphique de représentation des systèmes physiques qui fait partie de la classe des représentations de type réseau, la notion du bond-graph est une étape d'abstraction supplémentaire du graphe linéaire par laquelle l'unification des représentations physiques a été achevée. Elle a de plus l'avantage de permettre la représentation aisée et explicite des relations physiques mutuelles de certaines relations topologiques non représentables par un graphe linéaire.

### 2.1 Représentation des transferts de puissance :

La modélisation bond-graph suppose qu'il est possible de séparer et de localiser les propriétés des objets physiques et donc de décrire ces objets comme un système de propriétés élémentaires reliées entre elles, le système physique est alors représenté par des éléments multi-ports (**nœuds**), interconnectés par des liens.

### 2.2 Les variables mises en jeu :

#### 2.2.1. Les variables de puissance (Variables Conjuguées):

Dans le cas général, tout lien de puissance porte deux variables duales appelées **effort** et **flux**. La puissance portée par le bond est alors le produit de ces deux variables.

$$\begin{array}{c} e \\ \hline f \end{array} \rightarrow P = e * f$$

Le sens de la demi-flèche indique le sens conventionnel de l'échange d'énergie. Ce postulat est appelé le **postulat de puissance** qui est vérifié à posteriori dans tous les domaines de la physique.

#### 2.2.2. Les variables d'énergie (Variables généralisées):

Les variables généralisées sont introduites pour décrire l'état du système et sont définies par :

- \* Le déplacement généralisé :  $q = \int f dt$
- \* L'impulsion généralisée :  $p = \int e dt$

### 2.4. Les Eléments Bond-graph :

Le bond-graph est constitué d'un ensemble d'éléments et de jonctions reliées entre eux par des branches qui sont :

#### 2.4.1. Des Eléments Actifs (les Sources) :

Ces éléments fournissent de la puissance au système, ils sont dits **actifs**. On distingue deux types de source :

- Les sources d'effort  $S_e$ ,
- Les sources de flux  $S_f$ .

#### 2.4.2. Des Eléments Passifs :

Ce sont des éléments dissipateurs de puissance.

- A. L'élément R : l'élément R est utilisé pour modéliser tout phénomène liant l'effort au flux. Il est caractérisé par la loi générique :  $\Phi_R(e, f) = 0$ . Dans le cas linéaire :  $e(t) = R f(t)$
- B. L'élément C : l'élément C est utilisé pour modéliser tout phénomène liant l'effort au déplacement. Il est caractérisé par la loi générique :  $\Phi_C(e, q) = 0$ .

$$\begin{array}{c} e \\ \hline f \end{array} \rightarrow C : C_1 \quad e = (1/C_1) \int f dt$$

- C. Elément I : l'élément I est utilisé pour modéliser tout phénomène liant l'effort au moment. Il est caractérisé par la loi générique :  $\Phi_I(p, f) = 0$ .

$$\begin{array}{c} e \\ \longrightarrow \\ f \end{array} \quad \text{I: } I_1 \quad f = (1/I_1) \int e \, dt$$

### 2.4.3. Des éléments de Jonction :

Les éléments (**0**, **1**, **TF**, **GY**) servent à coupler les éléments R, C, I, et composent la structure de jonction du modèle correspondant à l'architecture du système étudié, ils sont conservatifs de puissance.

#### A. La Jonction 0 (Jonction Parallèle) :

La jonction 0 sert à associer les éléments soumis au même effort. Elle est caractérisée par les équations :

$$\begin{aligned} e_1 &= e_2 = \dots = e_n \\ \sum \alpha_i f_i &= 0, \alpha_i = \pm 1 \end{aligned}$$

#### B. La Jonction 1 (Jonction Série) :

La jonction 1 sert à associer les éléments soumis au même flux. Elle est duale de la jonction 0 et est caractérisée par les équations :

$$\begin{aligned} f_1 &= f_2 = \dots = f_n \\ \sum \alpha_i e_i &= 0, \alpha_i = \pm 1 \end{aligned}$$

Remarque : - Le but de connecter les éléments par les jonctions est de construire des systèmes complexes en composant des éléments simples.

- Les bonds qui lient une jonction à un élément sont dits **externes**, et ceux qui joignent deux jonctions sont appelés **internes**.

#### C. Le Transformateur TF : L'élément TF est un élément 2-ports. Il représente un couplage inter-domaines ses équations sont :

$$\begin{array}{c} e_1 \\ \searrow \\ f_1 \end{array} \quad \text{TF} \quad \begin{array}{c} e_2 \\ \searrow \\ f_2 \end{array} \quad \left\{ \begin{array}{l} e_1 = m e_2 \\ f_2 = m f_1 \end{array} \right.$$

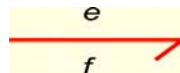
Si m n'est pas constant, on parle de transformateur modulé noté **MTF**.

#### D. Le Gyrateur GY : L'élément GY est un élément 2-ports. Il représente les phénomènes de couplage entre les différents domaines de la physique ses équations constitutives sont :

$$\begin{array}{c} e_1 \\ \searrow \\ f_1 \end{array} \quad \text{GY} \quad \begin{array}{c} e_2 \\ \searrow \\ f_2 \end{array} \quad \left\{ \begin{array}{l} e_1 = r f_2 \\ e_2 = r f_1 \end{array} \right.$$

### 2.5. Notion de Causalité :

Le langage bond graph (BG) est remarquable par la simplicité de ses concepts de base et son applicabilité à de très nombreux domaines industriels. Il est fondé sur le transfert d'énergie entre éléments, symbolisé par un lien de puissance



La puissance transférée est alors le produit entre les variables effort et flux :  $P = e * f$ . La dernière équation ci-dessus montre que, pour une puissance donnée, si la variable effort est imposée, alors la variable flux est automatiquement déterminée et réciproquement. Pour chaque lien BG, si l'information sur l'effort est imposée par l'élément à l'une des extrémités du lien, alors l'information sur le flux doit venir de l'autre extrémité du lien et vice-versa. Ce dernier point fait apparaître la notion de causalité. Supposons qu'on ait deux systèmes couplés A et B, tels que A transmet à B la puissance  $P=e \times f$ , nous avons deux situations possibles :

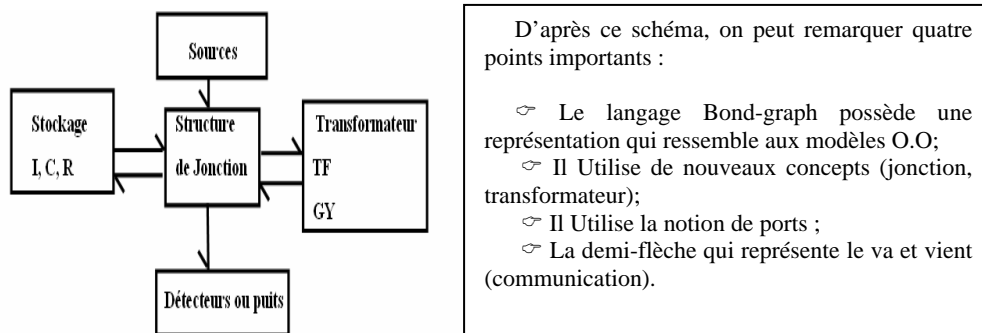
- A applique à B un effort e, qui réagit en envoyant à A un flux f ;
- A applique à B un flux f, qui réagit en envoyant à A un effort e.

Pour prendre en considération ces relations de cause à effet et les représenter sur le modèle bond-graph, nous introduisons le trait causal qui indique le sens dans lequel l'effort est connu.



### 3. Bond-Graphs et S.I (BGSi) [1, 2,4]:

On peut résumer la section 2 dans le schéma suivant :



Pour pouvoir adapter le langage Bond-graph dans la modélisation des systèmes d'informations, nous posons deux principes :

- ◆ **Principe 1** : L'information est une forme d'énergie dont les manifestations duales sont les données et les traitements (Effort=Traitements, flux=Données).
- ◆ **Principe 2** : L'échange d'information dans un système est assuré par un flux afin d'assurer la conservation totale d'information (système fermé).

Pour comprendre le principe 1, on a tiré un exemple concret qui fait partie des systèmes d'exploitation répartis (Migration de données, Migration de calculs, Migration de processus).

#### 3.1. Définition des Structures :

##### 3.1.1. Variables duales (Effort et Flux):

- **Effort** : D'après le principe 1, on a considéré que l'effort représente les traitements (appel des méthodes). Dans les méthodes orientées-objet l'exécution des méthodes se fait par l'envoi de messages. Un message est représenté sous la forme suivante :

- un receveur  $r$  (un objet),
- une méthode  $m$ ,
- des arguments éventuels  $a_1, a_2, \dots, a_n$ .

Dans notre cas, on peut considérer l'effort comme un message complexe (composé de plusieurs messages) et qui a la forme suivante :

- un ensemble de receveurs  $r_1, r_2, \dots, r_n$ . (des objets),
- des méthodes  $m_1, m_2, \dots, m_n$ ,
- pour chaque méthode  $m_i$  des arguments éventuels  $a_{i1}, a_{i2}, \dots, a_{in}$ .

- **Flux** : La structure des données est composée de deux parties :

- Données représentées par des structures,
- Etat (Retour : Booléen).

Remarque : Les structures des variables duales (Effort et flux) dépendent du système traité.

### 3.1.2. Structures des éléments:

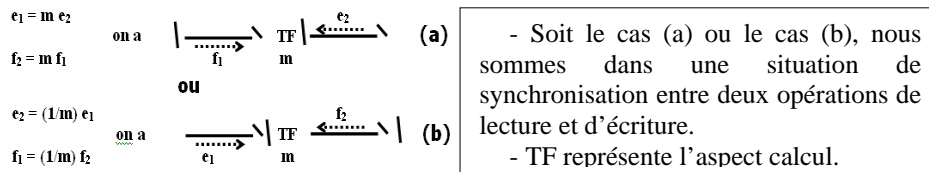
- \* La source d'effort représente les sources de traitements notées SE.
- \* La source de flux représente les sources de données notées SF.
- Ces deux objets représentent les objets actifs du système. (Exemple: Processus)

#### Objet de dissipation d'information R (Filtre):

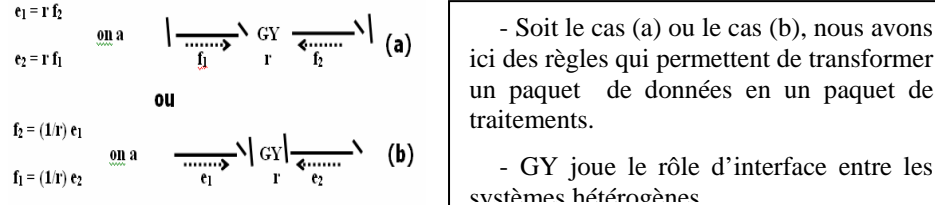
L'objet R (Filtre) peut jouer deux rôles :

- 1- Celui de filtre (flux entrant).
- 2- Celui de filtrage ou restriction (effort entrant).

#### Objet TF:



#### Objet GY:



#### Objet C:

Il représente l'élément d'accumulation d'information (exemple :BD).

#### Objet I:

Il représente l'évolution d'un objet dans le temps. (Ex: Fichier Logd'un O.S).

#### Les Objets Jonctions :

Ils jouent le rôle de relais ou de catalyseur d'information. Ce sont des objets complexes.

- o La Jonction 0 : (coopération)
  - ☞ Elle joue deux rôles:
    - \* Diffusion des messages,
    - \* Réception des données envoyées par les autres opérations.
  - ☞ Elle possède comme structure les files d'attente.
- o La jonction 1 : (concurrence)
  - ☞ Elle joue deux rôles:
    - \* Diffusion des messages,
    - \* Synchronisation des efforts.
  - ☞ Elle possède comme structure les files d'attente.

Remarque :

- ☞ Le nombre, les tailles et les types des files d'attentes représentent les paramètres des jonctions 0 et 1,
- ☞ Les files d'attentes utilisées dans la jonction 0 diffèrent de celles utilisées dans la jonction 1.
- ☞ Les jonctions et les transformateurs (GY, TF) utilisent des structures Indépendantes du système traité (Ex : Socket).

### 3.2 Démarche :

La théorie des bond-graphs pour les systèmes d'information stipule une démarche descendante qui consiste à localiser les rôles des objets, puis ascendante pour la construction du modèle conceptuel. La démarche que nous préconisons est la suivante :

- a - Identifier tous les objets du système à modéliser,
- b - Spécifier chaque objet, c-à-d définir sa partie statique et sa partie dynamique,
- c - Donner à chaque objet du système un rôle,
- d - Identifier et classer tous les événements du système à modéliser (décomposer les opérations complexes en opérations élémentaires).
- e - Trouver les relations de communications qui existent entre les objets pour chaque niveau de décomposition, c-à-d trouver les jonctions qui existent en commun.
- f - Si on a un système complexe alors il faut en définir les frontières.

## 4. BGSi et UML :

### 4.1 Définition du diagramme de classes :

En se basant sur l'utilisation du langage bond-graph dans les langages de simulation orienté-objet (modélica) [5, 6,7], on peut facilement remarquer deux choses :

La première est que le langage bond-graph n'est en sorte qu'une extension des méthodes orientées-objets avec cependant un certain nombre de différences. Tout d'abord les objets sont typés et possèdent un rôle évolutif, ensuite la communication entre eux ne se fait plus par envoi de messages mais par un envoi d'informations qui peut prendre les deux formes duales : le type donnée et le type traitement (effort-flux).

La deuxième est qu'on peut facilement déduire les classes du langage bond-graph dans les systèmes d'informations où on peut considérer un objet bond-graph comme un objet composé de plusieurs composants (Components). Ces composants peuvent être des objets source (source d'effort SE et source de flux SF), des objets de stockages (C, I), des objets de dissipation, des objets de transformation (transformateur TF, gyrateur GY) ou des objets de jonction (jonction 0 et jonction 1). Dans ce cas on a une relation de composition et une relation de spécialisation [1].

En se basant sur la notion de ports (interface), on peut aussi définir une autre hiérarchie de classes (interface Bondport, interface OnePortPassive, interface TwoPortPassive, interface OnePortEnergetic, etc. Chaque classe est une spécialisation d'une autre classe et peut avoir ses propres paramètres [7].

#### \* Port:

- ☞ Il permet de spécifier les points d'interaction d'un objet.
- ☞ Il indique les interfaces proposées par l'objet (realized) ainsi que les interfaces requises (required).
- ☞ Il existe deux types de ports :
  - Un port comportemental (Behaviour port)
  - Un port protocole (protocol port)



Tableau comparatif entre UML et BGSI [6]:

Modèle UML	Approche BGSI
* Objets Entités	* Objets I, C, R
* Objets Interfaces	* Ports
* Objets Contrôles	* Jonctions 0,1 ; Objets TF et GY
* Signal	* Bond
* Acteurs	* Objets Sources (Se, Sf)

## 5. Conclusion :

Cet article montre que les bond-graphs qui sont un outil puissant pour la modélisation des systèmes physiques peuvent être aussi adaptés pour la modélisation des schémas généraux de bases de données où l'avantage est de répondre aux préoccupations dynamiques de la communication ainsi que de la maîtrise du processus de modélisation qui se fait sur la base de la simulation. De plus la définition de la structure de classes dans ce langage pour les systèmes d'information peut être facilement générée en se basant sur des langages de modélisation et de simulation orientés-objet dans les systèmes physiques.

Ce travail est un premier pas, il existe plusieurs extensions et perspectives futures qui restent à envisager. Les plus importantes d'entre elles peuvent être résumées ainsi :

☞ Elever un peu le degré d'abstraction (étude théorique) en s'intéressant aux résultats intéressants des Bond-Graphs dans les systèmes physiques et dynamiques, par exemple

- \* La notion de contrôlabilité et d'observabilité,
- \* L'étude de la causalité,
- \* la définition de l'état d'un système, etc.

☞ Définir les règles de passage du flux de données en traitements et vice-versa qui se trouvent dans l'objet gyrateur GY.

☞ Modéliser des bases de données et de faire une étude comparative avec les autres méthodes de modélisation (les réseaux de Petri, etc.).

☞ Modéliser les bases de données Multi niveaux.

☞ Définir un langage formel pour la modélisation des protocoles de communications.

☞ Concevoir et modéliser des systèmes d'exploitation multi-tâches (système LAN).

Finalement, nous pensons que l'évolution de la théorie des BOND-GRAPH dans les systèmes d'information peut prendre la même allure que celle des réseaux de pétri (des réseaux ordinaires, réseaux à prédicats, réseaux colorés etc.).

## Références

- [1] Abed bahtou B., Rahmouni M.K. "Approche Bond-Graph pour la Modélisation des Systèmes d'Information (BGSI): Une Nouvelle Approche Orientée Communication" Mémoire De Magister, université Oran Es-Sénia, Février 2008.
- [2] Abed bahtou B., Rahmouni M.K., Haffaf H. "Modelling Dynamics In Information Systems By Using The Bond-Graphs Approach: A New Communication Oriented Approach" Séminaire international, Mai 1997, Computer Science and Information Systems Dept. Philadelphia University, Amman, Jordan.
- [3] Hafid Haffaf "Outil Bond-Graph pour la modélisation des systèmes physiques et Dynamique : nouvelle approche pour les systèmes d'information". Thèse de magister, Juin 1993.
- [4] Hafid Haffaf "Le concept d'état dans un B.G.S.I.". 1st regional seminar sur les Systèmes d'information", EX-IAP, université d'oran (es- sénia), Mars 1996.
- [5] Jan F. Broenink "Object-oriented modelling with bond graphs and Modelica" Simulation Series, Vol 31, No1, pp163-168, ISBN 1-56555-155-9, San Francisco, Janvier 1999.
- [6] Peter Fritzson & Vadim Engelson "Modelica - A Unified Object-Oriented Language for System Modeling and Simulation" E. Jul (Ed.): ECOOP'98, LNCS 1445, pp. 67-90, 1998. c Springer-Verlag Heidelberg Berlin 1998.



## Proposition d'une Méthode pour Améliorer la Qualité d'une Conception Orientée Objet

Lilia Grati<sup>1</sup>, Mohamed Tmar<sup>1</sup> et Faïez Gargouri<sup>1</sup>

<sup>1</sup>Laboratoire MIRACL, Institut Supérieur d'Informatique et de Multimédia de Sfax,  
BP 242 - 3021, Sakiet Ezzit, Sfax, Tunisie.

{lilia.grati, mohamed.tmar} @isimsf.rnu.tn, faiez.gargouri@fsegs.rnu.tn

**Résumé.** La qualité d'un logiciel résulte des décisions prises lors de la conception. C'est que la garantie d'une « bonne » qualité de modèles conceptuels, est vitale pour réduire le coût du logiciel. Dans cet article, nous présentons, en premier lieu, deux critères d'évaluation de la qualité d'une conception à savoir : les mesures et les heuristiques. En effet, les mesures sont considérées comme un indicateur permettant de prévoir la qualité d'un modèle conceptuel. Cependant, nous voyons que la seule application de ce critère est insuffisante pour juger la qualité d'une conception. De ce fait, il s'avère nécessaire de le combiner aux heuristiques afin d'améliorer les décisions conceptuelles du développeur. Nos contributions consistent d'abord, à associer ces heuristiques au contexte du processus unifié. Ensuite, nous exprimons ces dernières dans le langage formel : Z. Enfin, nous proposons un critère permettant d'améliorer la qualité des diagrammes d'interaction d'UML.

**Mots clés:** Qualité, conception orientée objet, mesures, heuristiques, processus unifié et langage Z.

### 1 Introduction

La qualité est un facteur essentiel de la pérennité des logiciels. Certes, la concurrence accrue entre les organismes spécialisés en génie logiciel a entraîné des attentes de plus en plus contraignantes en ce qui concerne la qualité des logiciels. Pour être compétitifs, il convient de réfléchir à l'amélioration de la qualité du produit logiciel dès les premières phases de son cycle de vie. Il est certain que la phase de conception a une importance cruciale dans le cycle de vie d'un produit puisqu'elle a un impact direct sur le coût et la qualité des phases qui la suivent. D'où, la corrélation entre le but d'avoir une bonne conception et le but de réduire la complexité des applications justifie les intentions des chercheurs à proposer divers critères pour évaluer la qualité d'une conception. Cet article est organisé comme suit :

Dans la deuxième section, nous présentons deux critères d'évaluation de la qualité d'une conception à savoir : les mesures et les heuristiques de conception. Ensuite, nous proposons d'étudier l'applicabilité des heuristiques de conception dans le contexte du processus unifié. Ceci permet d'orienter le concepteur à mieux utiliser ces heuristiques et à améliorer la qualité de ses décisions conceptuelles. La quatrième section consiste à exprimer ces heuristiques dans le langage formel Z afin d'enlever les ambiguïtés qui y sont détectées. Enfin, la cinquième section est consacrée à présenter un critère permettant l'amélioration de la qualité des diagrammes d'interaction d'UML.

### 2 Présentation de Deux Critères de Qualité pour évaluer une Conception

#### 2.1 Mesurage des Aspects Conceptuels

Fenton [1] définit le mesurage par *le processus d'assignation de nombres ou de symboles aux attributs des entités appartenant au monde réel afin de les décrire selon des règles bien claires et définies.*

Avant l'émergence de l'approche objet, les mesures traditionnelles qui ont été proposées sont jugées insuffisantes pour être utilisées dans le cadre de l'approche objet puisqu'elles ne s'adaptent pas aux concepts apportés par cette dernière. Nous présentons, dans ce qui suit, un survol sur un exemple de travaux basés sur les mesures de conception et de développement objet.

▪ **Chen et Lu**

Chen et Lu [2] ont proposé un ensemble de mesures, parmi lesquelles, nous présentons celles qui s'intéressent à l'aspect : complexité des classes.

- **OpCom**: la complexité des opérations d'une classe est définie par :

$$OpCom = \sum_{i=1}^m O(i) \quad (1)$$

$O(i)$  est la valeur de la complexité d'une opération  $i$  de la classe (avec  $i=1, \dots, m$ ), déduite à partir du tableau 1 présenté ci-dessous.

**Tableau 1.** Valeur de la complexité d'une opération [2]

Niveau de complexité	Valeur de complexité
Null	0
Très faible	1 – 10
Faible	11 – 20
Nominal	21 – 40
Elevé	41 – 60
Très élevé	61 – 80
Extrêmement élevé	81 – 100

- **OAC**: la complexité des arguments des opérations d'une classe est définie par :

$$OAC = \sum_{i=1}^m P(i) \quad (2)$$

$P(i)$  est la valeur de la complexité de chaque argument de l'opération  $i$  de la classe (avec  $i=1, \dots, m$ ). La valeur de  $P(i)$  est déduite à partir du tableau suivant.

**Tableau 2.** Valeurs des Arguments / Attributs [2]

Type	Valeur
Boolean ou integer	1
Char	1
Real	2
Array	3 – 4
Pointer	5
Record, Struct ou Object	6 – 9
File	10

- **AC** : la complexité des attributs de la classe est la somme des valeurs de tous les attributs de la classe.

$$AC = \sum_{i=1}^n R(i) \quad (3)$$

$R(i)$  est la valeur de chaque attribut  $i$  de la classe (avec  $i=1, \dots, n$ ), déduite à partir du tableau 2 présenté ci-dessus.

Les deux tableaux précédents ont fait l'objet de diverses critiques. A titre d'exemple, Henderson-Sellers [3] a jugé que les mesures utilisant des pondérations subjectives et arbitraires n'avaient pas de validité scientifique. D'où, les niveaux estimatifs des complexités des opérations ('très faible', 'faible', etc.) dans le tableau 1 ou les valeurs choisies pour les attributs et les arguments dans le tableau 2 n'ont pas de base théorique déterminée.

## 2.2 Heuristiques de Conception

Les heuristiques de conception sont considérées comme un critère de qualité aussi important que les mesures.

Selon K. Erni [4], *une heuristique de conception est une déclaration ou une information distillée d'une expérience de travail, basée sur les méthodologies de développement de logiciel et permettant d'améliorer la qualité de la conception.*

Dans ce contexte, nous avons étudié les travaux de Quatrani [5] et de Riel [6] qui a proposé une diversité d'heuristiques s'appliquant à sept axes. Nous allons présenter une sélection de ces heuristiques dans la section suivante.

### 3 Applicabilité des Heuristiques dans le Processus Unifié

Selon Jacobson et al. [7], si UML est un moyen standard de spécifier et communiquer les artefacts d'un système logiciel, il n'offre pas de méthodes à suivre et doit être alors utilisé dans le cadre d'un processus complet. D'où, nous proposons d'étudier l'applicabilité des heuristiques de conception étudiées dans le cadre d'un processus générique de développement de logiciel tel que le processus unifié. Notre objectif est d'aider le concepteur à savoir exactement quelles heuristiques adéquates à appliquer tout au long de son passage par les étapes de développement de son système.

#### 3.1 Cycle de Vie du Processus Unifié (UP)

UP s'articule autour de quatre phases : création, élaboration, construction et transition. Chacune de ces phases met l'accent sur une ou plusieurs activités à savoir : l'expression des besoins, l'analyse, la conception, l'implémentation et la transition.

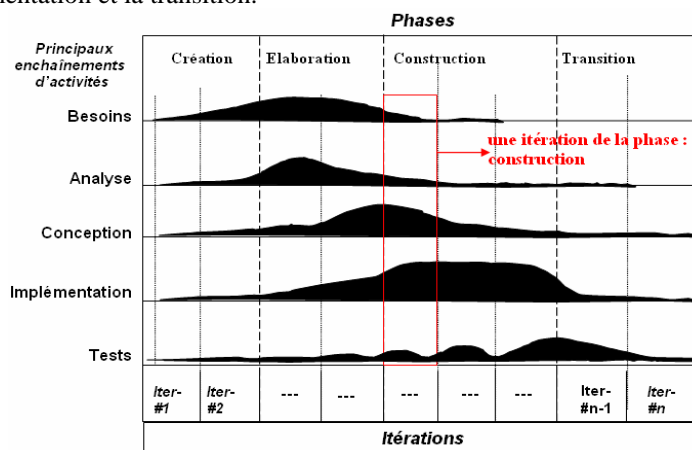


Fig. 1. Cycle de vie du processus unifié [7]

#### 3.2 Association des Heuristiques de Conception au Processus Unifié

Cette section est consacrée à présenter notre première contribution qui consiste à analyser l'applicabilité des heuristiques de Quatrani [5] et de Riel [6] dans UP.

##### ◇ Heuristique de Quatrani

**Heuristique 1 :** *Il ne faut distinguer entre deux acteurs que s'ils doivent interagir de façons différentes avec le système.*

Nous avons associé cette heuristique à la phase de création et spécialement à l'activité d'expression des besoins car elles permettent de dégager les fonctionnalités fondamentales remplies par le système et activées par un ensemble d'acteurs. Ainsi, cette heuristique peut être bénéfique pour prendre les précautions nécessaires dans le choix des acteurs. En outre, étant donné qu'UP est incrémental, on peut appliquer l'heuristique à la phase d'élaboration lors de la même activité afin de déterminer de nouveaux cas d'utilisation et acteurs.

##### ◇ Heuristique de Riel

**Heuristique 4.3 :** *Minimiser le degré de collaboration entre les classes, c'est-à-dire réduire le nombre de messages envoyés d'une classe à ses collaboratrices.*

On peut étudier l'applicabilité de cette heuristique à la phase d'élaboration qui reflète les interdépendances entre les classes. Ainsi, le couplage entre les classes, à travers l'échange de messages entre les classes, est modélisé par les diagrammes d'interaction (de collaboration et de séquence). Donc, cette heuristique peut être associée aux activités d'analyse et de conception permettant de produire respectivement ces deux diagrammes. En outre, elle peut être appliquée lors de la phase de construction. En particulier, durant l'activité d'implémentation, le degré de couplage d'une classe est exprimé en précisant les interactions entre les objets à travers les invocations de méthodes.

#### 4 Formalisation des Heuristiques dans le Langage Formel : Z

Notre deuxième contribution consiste à exprimer des heuristiques étudiées, dans le langage formel : Z, afin de vérifier la syntaxe de leurs spécifications et d'éliminer les ambiguïtés qui y sont détectées. Dans ce qui suit, nous présentons un exemple illustratif d'heuristiques traduites en Z. notons que d'autres formalisations d'heuristiques existent dans [8].

##### ◇ Heuristique de Quatrani

**Heuristique 2:** *Eviter d'avoir des classes redondantes.*

Deux classes sont redondantes si leurs attributs et opérations sont les mêmes.

$  \begin{aligned}  & \text{Synonym: } \text{Class} \rightarrow \text{F Class} \\  & \forall c1, c2: \text{Class} \\  & c2 \in \text{Synonym } c1 \\  & \Leftrightarrow c1.\text{name}.\text{syntax} \neq c2.\text{name}.\text{syntax} \\  & \wedge ((\forall x: c1.\text{attributes} \\  & \quad \cdot (\exists y: c2.\text{attributes} \\  & \quad \quad \cdot (x.\text{name}.\text{syntax} = y.\text{name}.\text{syntax} \\  & \quad \quad \vee x.\text{name}.\text{semantic} = y.\text{name}.\text{semantic}))) \\  & \wedge (\forall a: c1.\text{operations} \\  & \quad \cdot (\exists b: c2.\text{operations} \\  & \quad \quad \cdot (a.\text{name}.\text{syntax} = b.\text{name}.\text{syntax} \\  & \quad \quad \vee a.\text{name}.\text{semantic} = b.\text{name}.\text{semantic})))) \\  & \vee ((\forall x: c2.\text{attributes} \\  & \quad \cdot (\exists y: c1.\text{attributes} \\  & \quad \quad \cdot (x.\text{name}.\text{syntax} = y.\text{name}.\text{syntax} \\  & \quad \quad \vee x.\text{name}.\text{semantic} = y.\text{name}.\text{semantic}))) \\  & \wedge (\forall a: c2.\text{operations} \\  & \quad \cdot (\exists b: c1.\text{operations} \\  & \quad \quad \cdot (a.\text{name}.\text{syntax} = b.\text{name}.\text{syntax} \\  & \quad \quad \vee a.\text{name}.\text{semantic} = b.\text{name}.\text{semantic})))) \\  & \forall c1, c2: \text{Class} \cdot c1 \in \text{Synonym } c2 \Leftrightarrow c2 \in \text{Synonym } c1 \\  & \forall c1, c2, c3: \text{Class} \\  & \cdot c1 \in \text{Synonym } c2 \wedge c2 \in \text{Synonym } c3 \Rightarrow c1 \in \text{Synonym } c3  \end{aligned}  $
---

**Fig. 2.** Formalisation de l'heuristique 2 de Quatrani

#### 5 Minimisation du Couplage pour l'amélioration de la Qualité des Diagrammes d'interaction

Les diagrammes d'interaction (de collaboration et de séquence) correspondent aux scénarios des cas d'utilisation reflétant les besoins des utilisateurs. Ainsi, il s'avère utile de définir des critères permettant d'assister le concepteur lors de la modélisation de ces diagrammes et de le guider pour mieux représenter ces besoins. Pour ce faire, nous avons mené une étude empirique et statistique en modélisant deux corpus de diagrammes de séquence respectivement de « bonne » et de « mauvaise » qualité et ce, afin de déterminer les valeurs adéquates des critères proposés.

### 5.1 Problématique et Objectifs à Atteindre

En général, la qualité d'un diagramme de séquence (ou de collaboration) peut être considérée « bonne » s'il reflète la réalisation d'un scénario décrit pour un cas d'utilisation donné. Cependant, si nous nous référons à certaines heuristiques de conception étudiées dans notre état de l'art, telles que celles de Riel [6], nous remarquons une incitation à minimiser le couplage entre les objets en réduisant le nombre de messages échangés. Ceci a été de même constaté d'après l'étude des hypothèses et des mesures proposées pour la conception et le développement objet. En effet, une multitude de chercheurs tels que Chidamber et al. [9] [10], Li et al. [11] et Hitz et al. [12], ont confirmé théoriquement et empiriquement que le couplage excessif entre les classes est un facteur nuisible dans la conception modulaire. Plus le degré de dépendance entre les classes croît, plus leur réutilisation sera difficile puisque la sensibilité aux changements des parties du modèle est élevée.

Dans ce qui suit, nous proposons un critère permettant d'améliorer la qualité de la représentation des messages dans les diagrammes d'interaction et qui correspond à la vérification du nombre de messages qui y sont échangés.

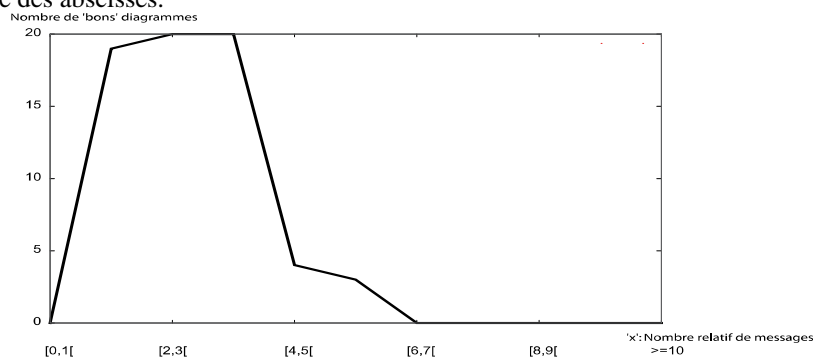
### 5.2 Vérification du Nombre de Messages Echangés entre les Objets

Certes, la seule minimisation du nombre de messages dans un diagramme d'interaction, ne peut pas être considérée comme un critère pertinent de qualité dans ce type de diagrammes. A titre indicatif, un diagramme d'interaction contenant un seul message ne peut pas être jugé de « bonne » qualité. De même, un diagramme contenant un grand nombre d'objets et ayant un nombre très réduit de messages, serait certainement de « mauvaise » qualité. Compte tenu de ce qui précède, nous pouvons assumer que le nombre de messages et le nombre d'objets sont deux critères interdépendants. En conséquence, il s'avère intéressant de montrer cette corrélation à travers la représentation de la variation du nombre de messages en fonction du nombre d'objets contenus dans les diagrammes de séquence modélisés dans les deux corpus.

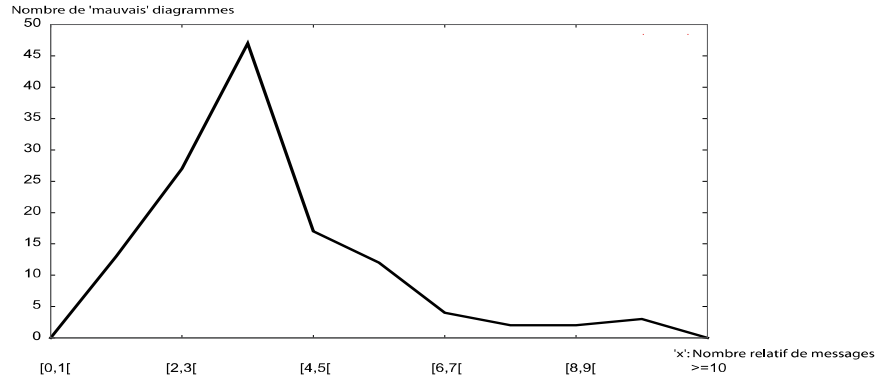
Etant donné un diagramme de séquence contenant 'No' objets et 'Nm' messages, nous définissons le nombre relatif de messages : 'x' par cette formule :

$$x = \frac{Nm}{No} \quad (4)$$

Sur la base de calculs effectués pour 66 diagrammes de « bonne » qualité et 127 diagrammes de « mauvaise » qualité, nous avons obtenu les deux courbes suivantes. L'axe des ordonnées contient les nombres de diagrammes ayant une valeur du nombre relatif de messages : 'x' appartenant à un intervalle donné dans l'axe des abscisses.



**Fig. 3.** Variation du nombre de diagrammes de « bonne » qualité en fonction de 'x'



**Fig. 4.** Variation du nombre de diagrammes de « mauvaise » qualité en fonction de 'x'

A partir des deux courbes précédentes, nous pouvons déduire l'existence d'une différence entre les valeurs des abscisses auxquelles correspond le maximum de chaque courbe. En effet, cette différence implique la possibilité de considérer le nombre relatif de messages comme un critère de qualité dans les diagrammes d'interaction. De plus, nous pouvons affirmer que les deux courbes suivent la même loi statistique à savoir : la *loi normale*. La fonction associée à cette loi est donnée par :

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2} \quad (5)$$

Où:  $\mu$  : moyenne des valeurs correspondant au nombre relatif de messages : 'x'.

$\sigma$  : écart type des valeurs du nombre : 'x'.

Afin de déterminer les valeurs adéquates correspondant au nombre de messages échangés dans les diagrammes d'interaction ayant un nombre d'objets donné, nous proposons de chercher d'abord la valeur de: ' $\mu_i$ '. Il s'agit en fait, de la moyenne des messages échangés dans les diagrammes de séquence contenant ' $i$ ' objets.

$$\mu_i = \frac{\sum_{j=1}^n nm_j}{n} \quad (6)$$

$nm_j$ : nombre de messages échangés dans un diagramme de séquence contenant ' $i$ ' objets;

$n$ : nombre de diagrammes de séquence modélisés, contenant ' $i$ ' objets;

En deuxième lieu, nous proposons de calculer l'écart type:  $\delta_i$  afin de déterminer l'intervalle de variation de la moyenne : ' $\mu_i$ '.

$$\delta_i = \frac{1}{n} \sqrt{\sum_{j=1}^n (nm_j - \mu_i)^2} \quad (7)$$

Donc, si on souhaite estimer le nombre adéquat de messages échangés dans un diagramme d'interaction, nous nous basons sur les valeurs prises par ces deux variables et que nous avons déterminées suite à notre étude empirique. Il s'agit alors, de vérifier si le nombre de messages échangés varie dans l'intervalle  $[\mu_i - \delta, \mu_i + \delta]$ . Ce dernier est considéré comme un intervalle de confiance.

Par exemple, pour un diagramme de séquence contenant 2 objets, nous avons déterminé :

$$\left. \begin{array}{l} \mu_2 = 4,57 \\ \delta_2 = 0,24 \end{array} \right\} \begin{array}{|l} \text{Dans un diagramme de "bonne" qualité contenant 2 objets, le} \\ \text{nombre adéquat de messages échangés est égal à 5.} \end{array}$$

## 6 Conclusion et Perspectives

Dans cet article, nous avons présenté deux critères d'évaluation de la qualité d'une conception orientée objet : les mesures et les heuristiques. Dans la troisième section, nous avons analysé l'applicabilité d'une sélection des heuristiques de Quatrani [5] et de Riel [6] aux phases et aux activités du processus unifié et ce, afin d'aider le concepteur à améliorer ses prises de décisions lors de la modélisation conceptuelle de son système. En étudiant ces heuristiques, nous y avons détecté des ambiguïtés. Nous avons alors considéré qu'il est utile de les exprimer dans le langage formel Z permettant ainsi de vérifier leurs syntaxes. Finalement, l'intérêt de la cinquième section était de proposer un critère permettant de minimiser le couplage dans les diagrammes d'interaction afin d'y améliorer la représentation des messages.

Les perspectives de notre travail consistent en premier lieu, à proposer d'autres critères de qualité s'appliquant particulièrement aux diagrammes d'interaction.

Notre intérêt portera aussi à étudier la qualité de la conception d'autres diagrammes UML en proposant de nouvelles heuristiques et mesures.

## Références

1. Fenton N.E.: Software Measurement: A Necessary Scientific Base. IEEE Transactions on Software Engineering, vol. 20. N° 3 (1994) 262 – 266.
2. Chen J.Y. and Lu J.F.: A New Metric for Object-Oriented Design. Information and Software Technology, vol. 35. N° 4 (1993) 232 – 240.
3. Henderson-Sellers B.: Object-Oriented Metrics: Measures of Complexity. Prentice-Hall (1996).
4. Erni K. and Lwerentz C.: Applying Design-Metrics to Object-Oriented Frameworks. In Proceedings of the 3rd International Software Metrics (1993).
5. Quatrani T.: Modélisation UML avec Rational Rose. Edition Eyrolles (2000).
6. Riel A.J.: Object Oriented Design Heuristics. Addison-Wesley Publishing Company (1996).
7. Jacobson I., Booch G., Rumbaugh J. : Le Processus unifié de développement logiciel. Editions Eyrolles (2000).
8. Grati L., Tmar M. and Gargouri F.: Using the Heuristics to Improve the Object-Oriented Design Quality. In Proceedings of IADIS Information Systems, International Conference (2008).
9. Chidamber S.R., Kemerer C.F.: A Metrics Suite for Object Oriented Design. IEEE Transactions on Software Engineering, vol. 20. N° 6 (1994) 476 – 493.
10. Chidamber S.R., Darcy D., Kemerer C.F.: Managerial use of Metrics for Object Oriented Software: An Exploratory Analysis. IEEE Transactions on Software Engineering, vol. 24. N° 8 (1998) 629 – 639.
11. Li W., Henry S., Kafura D., Schulman R.: Measuring Object-Oriented Design. Journal of Object-Oriented Programming, vol. 8. N° 4 (1995) 48 – 55.
12. Hitz M. and Montazeri B.: Measuring Coupling and Cohesion in Object-Oriented Systems. Proc. Int'l Symposium on Applied Corporate Computing, Monterrey, Mexico (1995).

## *Geographic information Systems*

---





## Vers un système d'unification des systèmes d'informations géographiques

ABDICHE Fethi et ATMANI Baghdad

Département d'Informatique, Faculté des Sciences,  
Université d'Oran Es-Senia,  
BP 1524, El-M'Naouer, 31000, Oran, Algérie

[fethi\\_abdiche\\_box@yahoo.fr](mailto:fethi_abdiche_box@yahoo.fr), [atmani.baghdad@univ-oran.dz](mailto:atmani.baghdad@univ-oran.dz)

**Résumé.** Les systèmes d'informations géographiques (SIG) ont été conçus et développés à l'origine pour des besoins bien déterminés. Ces systèmes sont donc hétérogènes, autonomes et indépendants. Par conséquent, les formats de stockage des données spatiales et descriptives fournies par ces différents SIG sont fortement hétérogènes. Nous nous retrouvons donc confrontés à un problème d'unification des bases de données géographiques. Nous proposons dans cet article un modèle d'intégration des données géographiques à références spatiales hétérogènes et indépendantes. L'objectif de cette phase d'intégration est la préparation des bases de données géographiques pour l'analyse et l'extraction des connaissances à partir des données structurées. Cet article est structuré comme suit. La section 2 trace l'évolution de la notion des données géographiques et leurs représentations depuis leurs formats vectoriels jusqu'au format relationnel au sein d'un SGBDR. La notion de liaison logique entre les cartes géographiques représentant le même contexte géographique est présentée dans la section 3. La problématique du système d'unification proposé est abordée dans la section 4. Nos résultats et quelques travaux futurs seront donnés en guise de conclusion.

**Mots clés :** Système d'informations géographiques, bases de données à références spatiales, Oracle spatial, intégration de données hétérogènes

### 1 Introduction

Nous voyons notre monde quotidien changer de plus en plus vite, et l'information, qui en est l'image instantanée, devient de plus en plus difficile à assimiler directement. Sa représentation classique, sous les formes longtemps satisfaisantes, est maintenant peu adaptée. Elles nécessitent de connaître les informations sur les besoins sanitaires, économiques et écologiques.

Les systèmes d'information géographique (SIG) sont une réponse à ces impératifs dans différents domaines. Ainsi, l'intégration de nouvelles sources de plusieurs natures, leurs combinaisons et leurs analyses permet de présenter l'information de façon efficace et directement exploitable.

Tout d'abord réservés aux professionnels de la géographie, les SIG ont vus rapidement leur champ d'application s'étendre dans tous les domaines où la prise de décision rapide en fonction des données spatiales est primordiale: environnement, gestion des réseaux (eau, électricité, téléphone...), aménagement urbain, marketing, prospection géologique... [5].

À l'origine, les systèmes d'information ont été conçus et développés pour les besoins spécifiques. Ces systèmes sont donc hétérogènes, autonomes et indépendants [10]. La commercialisation à grande échelle des SIG a provoqué une croissance très rapide en terme de production, stockage et diffusion de ces informations qui représentaient des phénomènes liés au territoire urbain concernant une population  $x$  au cours d'une période  $y$ . Ces phénomènes sont sauvegardés sous formes de cartes géographiques vectorielles ou en bases de données à références spatiales. Ces dernières sont très variées. De là, le besoin d'intégration de ces données est devenu une tâche primordiale pour une analyse complète du territoire. Cette analyse qui s'est limitée au calcul de la distribution de la population dans un périmètre bien défini ou à la réponse à de simples requêtes de ce genre, sans arriver à extraire les relations existantes, implicites qui nous aident à mieux comprendre notre territoire, cette extraction de relations implicites à partir des données à références spatiales est connue sous le nom de Fouille de données spatiale (FDS).

Le contexte de notre travail englobe deux notions indissociables : l'intégration des données à références spatiales et l'extraction de la connaissance à partir de ces masses de données [3,4] qui alimentent le processus de la FDS : population, scolarisation, santé, réseau routier, réseau d'électricité,

stockées sous divers formats. Or, ces données sont fortement hétérogènes. Donc notre objectif par le présent article est l'intégration des données spatiales en premier lieu.

## 2 Les Systèmes d'informations Géographiques

Le système d'information géographique (SIG) est un système de gestion de base de données pour la saisie, le stockage, l'extraction, l'interrogation, l'analyse et l'affichage de données localisées [6]. Le succès connu par les SIG témoigne le besoin de gérer des représentations spatiales. Les SIG sont de puissants outils pour manipuler, interroger et visualiser les bases des données spatiales. Une donnée spatiale est généralement composée de trois parties : une partie géométrique (la forme par exemple ligne, point, polygone), une partie descriptive (par exemple le nom, la population d'une ville) et une partie métrique calculée (par exemple la superficie, le périmètre) [12].

Nous nous intéressons à deux fonctions des SIG qui sont l'Acquisition et l'Archivage.

### 2.1 L'acquisition de l'information géographique

Les données peuvent provenir de fournisseurs extérieurs, de numérisation directe ou de traitement particulier, elles peuvent être fournies sous forme de bases de données relationnelles ou orienté objet, sous forme de documents XML ou GML et même mode vectoriel ou raster. Cette diversité peut être résumée en deux principales catégories d'informations : Informations descriptives ou Spatiales (géométriques).

La partie géométrique fournit des informations sur la position et la forme des objets étudiés. Cette information peut être sous forme vectorielle ou de raster [2] :

Le mode d'acquisition le plus répandu est la **couche**. En effet, les objets sont organisés en couches, chacune contient l'ensemble des objets homogènes (bâti, rivières, voiries, parcelles, etc ...) [1].

### 2.2 Archivage de l'information géométrique

Certains systèmes d'informations géographiques gèrent simultanément les données géométriques ainsi que les données descriptives alors que d'autres séparent ces deux types de données [9].

L'organisation et la maintenance des données en général sont des tâches complexes. L'ajout des composantes spatiales de plusieurs bases de données augmente encore cette complexité. L'intégration des données spatiales sous un SGBDR unique est apparue en cours des dix dernières années [7].

### 2.3 Unification de format de stockage des images vectorielles

La structure de l'image vectorielle est dans la plus part des SIG sous forme de plusieurs fichiers. Chaque SIG propose une structure à part de représentation de ces fichiers de stockage. L'accès et la manipulation des données dans ce mode nécessite un API propre au SIG. Nous citons par exemple *Spatial Database Engine (SDE API)* de ESRI, ou *Mapbasic API*, de MapInfo.

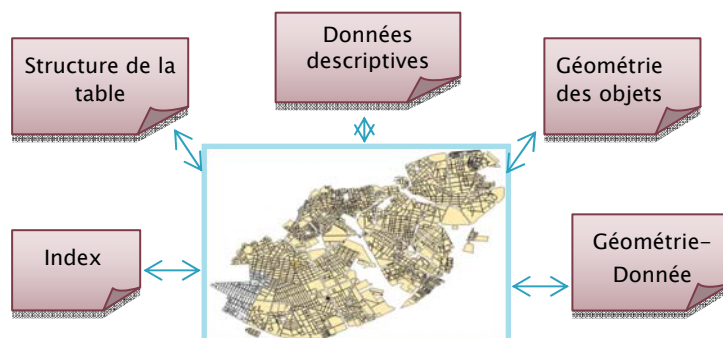


Fig. 1. La structure d'une image vectorielle sous MapInfo

Vu l'hétérogénéité des formats de stockages, notre objectif est d'importer les fichiers de description des images vectorielles sous un format unique, avec la création des liens logiques entre les différentes couches. Pour cela, nous procédons comme suit :

- La conversion des images vectorielles conçues sous différents SIG en un seul format (.tab) à l'aide de l'outil **Universal Translator** disponible sous MapInfo.
- Le stockage de chaque image vectorielle représentée par plusieurs fichiers (.tab) sous MapInfo vers la base de données Oracle sous forme de tables relationnelles. Cette opération permet de créer un champ de type **SDO\_GEOMETRY** pour le stockage des objets spatiaux.
- La création des liens logiques entre les tables relationnelles pour expliciter la propriété d'appariement entre les différentes couches géographiques (images vectorielles).

L'importance d'utilisation des tables relationnelles à références spatiales réside dans le fait que les objets géométriques ainsi que leurs données descriptives sont stockées dans la même table relationnelle d'un côté. Plusieurs organisations ont adopté Oracle Spatial comme leurs support de stockage de base, par la migration de leurs méthodes traditionnelles de maintenances des objets spatiaux vers un model relationnel.

Les langages de programmations comme java ou C++ sont utilisés pour créer, et extraire les données à partir d'une base de données Oracle [7].

### 3 Intégration des données spatiales hétérogènes

Nous avons constaté deux états de figures qui consistent à unifier les couches conçues sur des SIG hétérogènes. En premier lieu, les données géométriques sont stockées sous formes de fichiers de types différents, ou bien ils sont stockés dans des fichiers de même type, mais aucun lien logique n'existe entre les couches géographiques bien qu'elles représentent le même phénomène.

Une solution possible à ces problèmes passe par l'unification des bases de données, qui vise à rendre les relations plus explicites entre les divers éléments. Celle-ci peut prendre la forme d'une fusion [8].

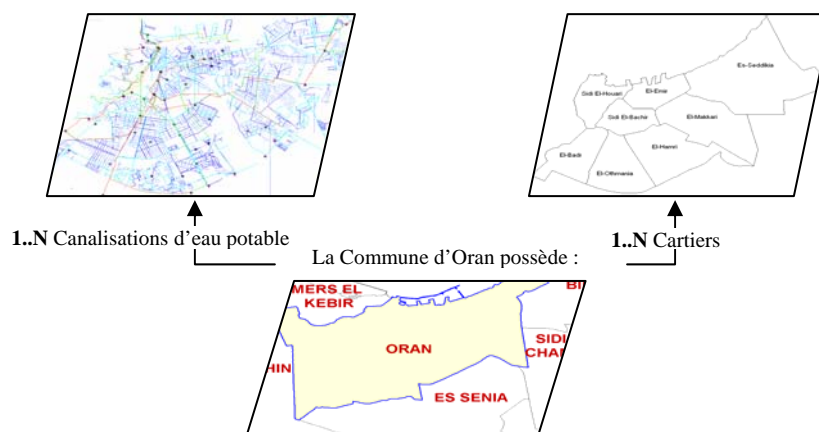


Fig. 2. Appariement des couches indépendantes

Dans cet axe, nous distinguons principalement le projet européen **MurMur** [8] (*Multi-representations and multiple resolutions in geographic databases*), qui consiste à concevoir un atelier de génie logiciel (AGL), consultable au laboratoire **COGIT**, qui permet de modéliser des schémas de données sur interface graphique. Le schéma conceptuel de la base de données unifiée est édité sur l'éditeur MADS [11] (*Model for Application Data with Spatio-Temporal Features*) développé dans le cadre du projet MurMur [8] sera en suite transformé en un schéma logique, automatiquement créé sur un SIG ou un SGBD cible. Les derniers développements du projet **COBALT** [13] (*Conception de Bases de données Localisées et Temporelles*) consistent à concevoir un module de translation **MADS2MI** qui permet la création d'un schéma logique cible, à partir de MADS vers le SIG MapInfo [13].

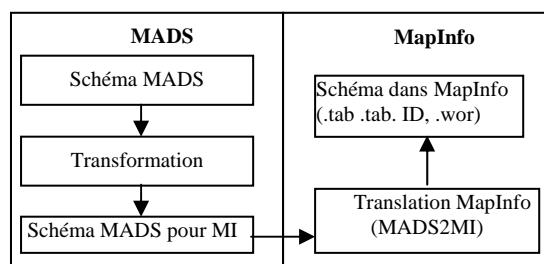


Fig. 3. L'application MADS2MI du projet Cobalt

Jusqu'ici nous avons explicité la problématique d'inexistence de lien entre les différentes couches géographiques au sein du même type de fichiers de stockage, la deuxième difficulté consiste à traiter des couches résidant dans des fichiers de stockage de types différents.

## 4 Contribution

La complémentarité entre les différentes sources d'informations géographiques hétérogènes est une réalité inévitable. La présence des différentes couches formant une seule carte et décrivant le territoire présente un support complet d'informations. Après avoir convertie les images vectorielles de différents formats en un seul sous *MapInfo* et sauvegarder le tout sous une base de données relationnelle d'Oracle, nous nous focalisation sur notre contribution qui consiste à expliciter l'appariement entre différents images vectorielles (couches) représentant le même contexte géographique. Notre objectif par la présente contribution consiste à créer un schéma relationnel intégrant les différentes couches géographiques indépendantes.

### 4.1 Modèle proposé

Durant les testes sur les cartes vectorielles converties en tables relationnelles géo référencées, ces dernières n'étaient pas dans certains cas, porteuses d'une information décrivant le territoire, donc nous supposons trois cas de figures :

- Les tables ne contenaient aucun champ représentant une clé primaire ou étrangère.
- Une des tables contient une clé.
- Les deux tables contiennent des clés.

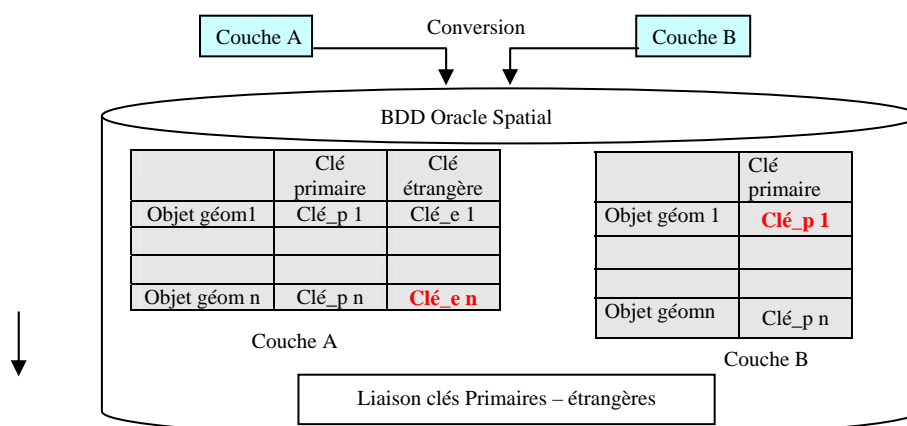


Fig. 4. Architecture du modèle proposé

Donc la difficulté réside dans la détection d'un lien logique sous forme de clé étrangère - clé primaire entre les tables géo référencées stockées sous Oracle Spatial.

Contrairement à la technique adoptée dans le projet MurMur [8], qui consistait à créer le schéma logique après avoir unifié les schémas. La technique utilisée ici, vise l'intégration des bases de données relationnelles géo référencées.

L'avantage de notre approche c'est d'éviter le développement d'outils d'interfaçages avec les SGBD ou les SIG, comme MADS2MI du projet **COBALT** [13].

## 4.2 Liaison clés Primaires – clés étrangères

Notre objectif dans le présent papier vise la mise en correspondance entre les différentes couches alors l'intégration des objets spatiaux s'effectue par géo localisation, c'est-à-dire « qui appartient à quoi ? ».

Afin de respecter cette modélisation, on a imaginé trois cas d'utilisations :

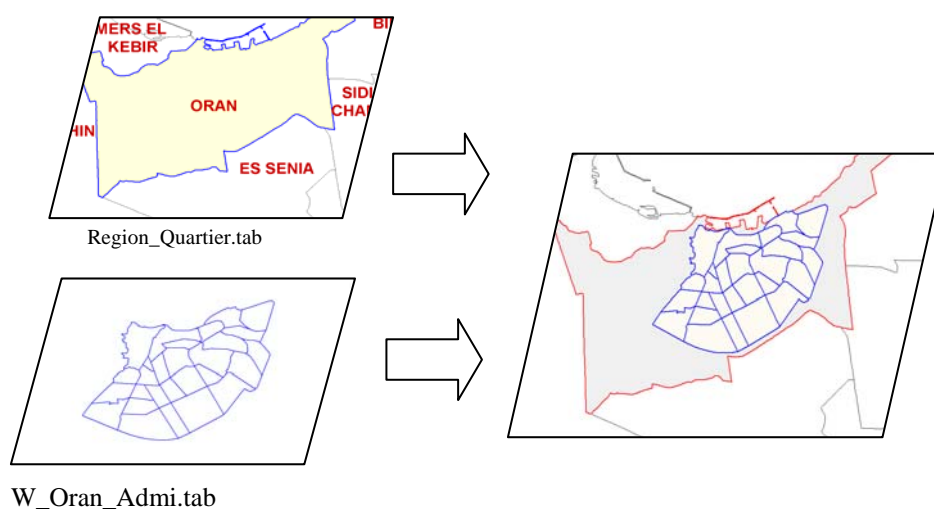
**Cas 1** : Les tables ne contenaient aucun champ représentant une clé primaire ou étrangère.

La solution est de créer deux nouveaux champs dans les deux tables, le premier correspond à la clé primaire et l'autre à la clé étrangère. La distinction se fait par l'expert

**Cas 2** : Une des tables contient une clé.

La même procédure du premier cas est appliquée, reste à définir la nouvelle clé (primaire ou étrangère) adéquate, dans l'une des tables.

**Cas 3** : les clés existent dans les deux tables, reste à créer le lien logique entre elles.



**Fig. 5.** Appariement entre deux bases de données géographiques

La création des clés s'effectue par le langage de manipulation de données **SQL**. L'expert identifie la clé primaire ou étrangère. Une interface graphique prendra en charge cette opération.

Pour le cas de la figure 5, on a procédé à l'intégration des deux cartes : **Region\_Quartier.tab**, fournie par le CNTS (*Centre National des Techniques Spatiales*) et **W\_Oran\_Admi.tab** par le département de la géographie de l'université d'Oran. Les deux cartes (couches) ont été conçues sous MapInfo, donc du même format d'échange mais avec deux schémas différents. Après leur conversion en tables relationnelles géo référencées sous Oracle Spatial, leurs structures étaient comme suit :

Region\_Quartier (Quartier, ID\_SEC\_HYD, MI\_PRINX)

W\_Oran\_Admi (Code\_Commune, Commune, Unité, Pole, MI\_PRINX)

Pour la table **W\_Oran\_Admi**, la clé primaire est : Code\_Commune ou Commune, il reste à définir la clé étrangère dans la table **Region\_Quartier** qui va référencer cette clé primaire, pour cela un nouveau champ sera créé et nommé Commune\_Ref.

- a. Suppression d'index sur w\_oran\_admi :  
`ALTER TABLE w_oran_admi DROP COLUMN MI_PRINX`
- b. Création d'une contrainte clé primaire sur la table référencée :  
`ALTER TABLE w_oran_admi  
ADD CONSTRAINT clé_primaire PRIMARY KEY (commune)`
- c. Rétablir l'index supprimé de la table w\_oran\_admi  
`CREATE INDEX MI_PRINX  
ON w_oran_admi (commune)`
- d. ajout d'une clé étrangère qui référence la clé primaire dans w\_oran\_admi  
`ALTER TABLE quartier_region  
ADD CONSTRAINT ref_to_commune FOREIGN KEY (commune_ref) REFERENCES  
W_oran_admi(commune)`
- e. Enfin, instancier la nouvelle clé étrangère dans la table des cartiers par le nom de la commune auquel ils font références :  
`UPDATE TABLE quartier_region  
SET commune_ref='ORAN'`

### 4.3 Application

L'intégration des sources de données géographiques par l'intervention d'un expert n'est en premier lieu pas pertinent. Autrement, qui a toute cette connaissance pour affecter des clés à des tables indépendantes et les référencé ?

Par conséquence, l'apport de cet article est de structurer les cartes vectorielles et les unifier sous une seule forme afin de faciliter d'autres processus d'intégration de données descriptives sous forme relationnelles ou XML dans un objectif d'avoir un entrepôt de données assez complet pour servir comme base de départ vers un processus de fouille de données [3] spatiale.

Une autre utilité de cette proposition c'est la facilité d'exécution des requêtes une fois les données spatiales structurées suivies d'intégration de données descriptives.

```
SELECT * FROM Table_Stations_service S
WHERE
S.commune IN (SELECT C.Nom_Commune
              FROM Commune C
              WHERE C.Taux_véhicule>100 000)
```

Dans cet exemple qui reste illustratif, toutes les stations services des communes dont le taux de véhicules est supérieur à 100 000 seront sélectionnées. Du moment où les stations services sont des composantes à références spatiales dans la table Table\_Stations\_service, elles peuvent être affiché sur une interface graphique, ce qui facilitera la décision en cas de problème d'aide à la décision.

## 5 Conclusion

Nous avons abordé dans cet article une problématique qui évolue à cause de l'évolution des structures des sources de données qui sont de plus en plus hétérogènes. Nous avons choisi de convertir les données spatiales en tables relationnelles sous Oracle pour unifier les différents formats d'échange et bénéficier de la facilité d'intégration des tables relationnelles. Cette structuration et unification des données spatiales et un pas vers d'autre processus d'intégration des données descriptives qui sont en relation avec les couches fournies et indépendantes des schémas logiques des tables géo référencées. Une fois les données spatiales et descriptives seront structurées et intégrées au sein d'un entrepôt de données géographiques, la phase suivante consiste à extraire des connaissances implicites [3,4].

## Références

1. Abdiche, F., Hamdadou, D. : Proposition d'un Modèle pour l'Exploitation des Bases de Données à Références Spatiales Distantes. Les troisièmes journées internationales sur l'information graphique, PP 229-234, Constantine (2007)
2. Armitage, T : Getting started with oracle spatial. Oracle Corporation, United States (2006)
3. Atmani, B., Beldjilali, B.: Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, Vol.26, N°2 (2007) 171-197.
4. Atmani, B., Beldjilali, B.: Neuro-IG : A Hybrid System for selection and elimination of predictor variables and non relevant individuals, INFORMATICA, International Journal, Vol. 18, N°. 2 (2007) 163-186.
5. Jérôme, F. : Le système d'information géographique de santé : un outil pour les soins de demain. Service d'Information et d'Informatique Médicales CHU de Grenoble, France (1994)
6. Laurini, R. : les base des données en géomatiques, Traité des nouvelles technologies, Hermé, Paris (1993)
7. Michael, D.: Integrating ArcGIS to Enterprise Oracle Spatial Using Direct Connect. Twenty Third Annual ESRI International User Conference GIS Serving Our World, San Diego, California, United States (2003)
8. Mustière, S., Gesbert, N., Sheeren, D.: Unification des Bases de Données Géographiques au Laboratoire COGIT. Bulletin d'information n.74 de l'IGN : Bilan de la recherche, pp.71-78, France (2002)
9. Marmonier, P. : L'information géographique, document de l'Ecole Nationale des Sciences Géographiques ENSG, France (2002)
10. Ouziri, M. : Utilisation des Topics Maps pour l'Interrogation et la Génération de Documents Virtuels. Thèse d'état présentée à l'institut national des sciences Appliquées de Lyon, France (2003)
11. Pinet, F. : Vers une Méthode de Conception pour la Télégéomatique : Intégration et Validation au sein d'un Atelier de Développement. Thèse d'état présenté à L'Institut National des Sciences Appliquées de Lyon, France (2002)
12. Rigaux, P., Scholl, M., Voisard, A.: Spatial databases with application to GIS. Morgan Kaufmann Publishers Inc, San Francisco, United States (2002)
13. Stéphane, J., Olivier, E. : Cobalt : Conception de Bases de Données Localisées Temporelles. Rapport final INTEREG II, Laboratoire THEMA, Université de Franche-Comté, Besançon et EPFL-LBD, Lausanne (2002)

## Utilisation d'un SIG pour l'évaluation des caractéristiques physiques d'un bassin versant et leurs influences sur l'écoulement des eaux (bassin versant d'Oued El Maleh, Nord-Ouest d'Algérie)

Bentekhici Nadjla<sup>1</sup>, Zebbar Zien ed dinne<sup>1</sup>

<sup>1</sup> Agence Spatiale Algérienne, centre des techniques Spatiales, BP 13 Arzew, Oran.  
[nadjla\\_b@yahoo.fr](mailto:nadjla_b@yahoo.fr)

**Résumé.** L'eau en tant que ressource vitale, sa protection nécessite une évaluation des paramètres physiques qui jouent un rôle important dans l'écoulement des eaux superficielles. Ces facteurs englobent les caractéristiques géologiques, géomorphologiques, hydrologiques, et climatiques. Nous nous intéresserons, dans la présente étude, aux caractéristiques morphométriques et hydrométriques du bassin versant en question. Les méthodes classiques utilisées dans les études des bassins versants sont basées essentiellement sur des techniques manuelles, avec l'apparition de nouveaux outils tels que les SIG et la télédétection, il est devenu aisé de déterminer les paramètres de forme, de relief et la typologie d'un réseau hydrographique ainsi qu'un gain de temps.

Ces nouvelles techniques ont permis de donner des informations numériques caractérisant le relief et la morphométrie du bassin, tout en se basant sur les modèles numériques de terrains. Elles sont utilisées pour estimer les risques naturels telles que les inondations. Les inondations présentent, par la gravité de leurs conséquences sur le plan humain et matériel, un risque naturel d'une grande importance. Pour faire face à ce risque, il est nécessaire de déterminer les caractéristiques d'un bassin versant. Ce travail a été effectué sur le bassin versant d'Oued El Maleh, qui est situé dans la région Nord-Ouest d'Algérie, il est d'une superficie de 787Km<sup>2</sup>.

**Mot clé :** bassin versant, caractéristiques physiques, SIG, MNT, écouleme.

### 1 Introduction

Les techniques classiques utilisées dans l'étude du complexe physique d'un bassin versant sont basées essentiellement sur des méthodes manuelles dont les résultats sont souvent imprécis, mais avec l'avènement des nouveaux outils tels que les SIG et la télédétection, il est devenu facile de déterminer les paramètres de forme, de relief et la typologie d'un réseau hydrographique.

L'utilisation de ces nouvelles techniques a permis ainsi de répondre aux exigences d'exploitation des eaux superficielles et la prévention contre les risques d'inondation.

### 2 Description de la zone d'étude et méthodologie de travail

Le bassin versant d'Oued El Maleh se situe dans la partie Nord-Ouest du territoire national. La zone d'étude est caractérisée par un climat méditerranéen aride à semi-aride avec des influences chaudes du Sahara au Sud et celles fraîches du Nord et de l'Est. Cette région est connue par une moyenne de précipitations de l'ordre de 300 mm/an. Le bassin versant d'étude se situe approximativement entre (1° 9' 24'' et 1°26' 17'' W) de longitude et entre (35°17'22'' et 35°16'37''N) de latitude.

Dans le cadre de cette étude, nous avons utilisé l'image optique LANDSAT ETM+ prise le 19- Mai 2002 d'une résolution de 30\*30m, ces images ont été utilisées pour l'actualisation du réseau hydrographique. Cette mise à jour consiste d'une part à analyser les paramètres d'un cours d'eau (forme, couleur, texture et structure) à partir d'une image satellite, d'autre part, à la correction de ce réseau par le chevelu digitalisé sur des cartes topographiques à 1/25.000.

Ces cartes seront utilisées aussi pour générer un modèle numérique du terrain.



L'extraction automatique des caractéristiques physiques d'un bassin versant sollicite l'utilisation des outils performants, dans le cadre de cette application, nous avons utilisé les logiciels MapInfo et Vertical Mapper.



Fig. 1. Localisation du bassin versant d'Oued El Maleh

### 3 Conception d'une base de données

L'opération consiste à élaborer d'une banque de donnée hydraulique, capable de décrire d'une manière plus poussée l'ensemble des paramètres qui agissent directement ou indirectement sur le fonctionnement du modèle physique (Bassin versant). L'ensemble de données géographique a été introduit dans le système d'information sous une forme numérique.

Le passage du monde réel au modèle conceptuel correspond à un processus de modélisation où les objets du monde réel sont classés en catégories et désignés par nom. Il consiste en trois éléments de base appelés entités, attributs et relation entre entités. Le dictionnaire des données contient toutes les informations sur les entités de base de données (le nom, la définition, référence spatiales, etc) (Dif A 1999) .

## 4. Résultat

### 4.1 Extraction automatique des caractéristiques du bassin versant

L'utilité de la connaissance de ces paramètres réside dans la caractérisation de l'environnement physique et dans l'information sur l'écoulement superficiel

#### 4.1.1 Les caractéristiques de forme :

Les paramètres de forme sont des éléments nécessaires pour caractériser l'environnement physique.

Nous avons, en premier lieu, délimité le bassin en question à partir des données satellitaires et du modèle numérique de terrain MNT.

En second lieu, nous avons déterminé la surface et le périmètre du bassin versant à partir d'une requête géographique en utilisant le langage SQL.

Table 1. Extraction automatique des: périmètre, surface et coefficient de compacité

Kc_oued_el_maleh Données			
Nom_bv	S_km2	P_km	Kc_
<input type="checkbox"/> O. el maleh	787.10	151.30	1.510

Partant d'une requête géographique et à partir de la surface et le périmètre du bassin, nous avons calculé le coefficient de **GRAVELIUS**.

Ceci implique qu'on est en présence d'un bassin d'une forme allongée, sachant que ce coefficient est proche de 1 pour un bassin de forme circulaire (Nouvelot, J-F 1989).

Un autre indice caractérisant la forme du bassin est le rectangle équivalent, les valeurs de longueur et de largeur ont été estimées dans le but de comparer les bassins entre eux du point de vue de l'influence de leurs caractéristiques géométriques sur l'écoulement.

**Table 2.** Extraction automatique de rectangle équivalent

rectangle_equivalent Données					
Nom_bv	S_km2	P_km	Kc	L_km	L_km_2
O. el maleh	787.10	151.30	1.510	47.26	28.35

#### 4.1.2 Les caractéristiques de relief

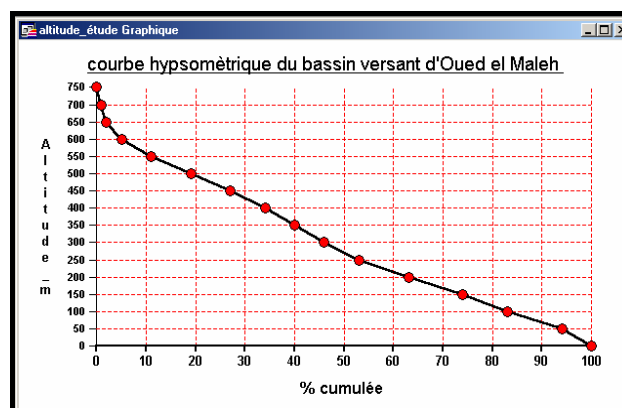
L'influence du relief sur l'écoulement se conçoit facilement, car de nombreux paramètres hydrométéorologiques varient avec l'altitude (précipitations, températures, etc.).

En outre, la pente influe sur la vitesse d'écoulement. Le relief se détermine lui aussi au moyen d'indices ou des caractéristiques suivantes

##### a) La courbe hypsométrique

La courbe hypsométrique fournit une vue synthétique de la pente du bassin, donc du relief.

L'extraction automatique du graphe hypsométrique a été faite à partir du modèle numérique d'altitude (MNA) et des requêtes géographiques.



**Fig. 2.** Courbe hypsométrique du bassin d'Oued El Maleh

Les courbes hypsométriques peuvent servir à la détermination de la pluie moyenne sur un bassin versant et donnent des indications sur le comportement hydrologique et hydraulique du bassin et de son système de drainage.

##### b) Caractéristiques des altitudes

L'altitude moyenne se déduit directement de la courbe hypsométrique. On peut la définir comme suit :

$$H_{moy} = \sum \frac{A_i \cdot h_i}{A} \quad (1)$$

$H_{moy}$  : altitude moyenne du bassin [m] ;  $A_i$  : aire comprise entre deux courbes de niveau [km<sup>2</sup>] ;  $h_i$  : altitude moyenne entre deux courbes de niveau [m] ;  $A$  : superficie totale du bassin versant [km<sup>2</sup>].  **$H_{moy} = 299\text{m}$ .**

L'altitude médiane correspond à l'altitude lue au point d'abscisse 50% de la surface totale du bassin, sur la courbe hypsométrique. Cette grandeur se rapproche de l'altitude moyenne, dans le cas où la courbe hypsométrique du bassin présente une pente régulière. L'altitude médiane dans le bassin d'Oued El Maleh est égale à **265 m**.

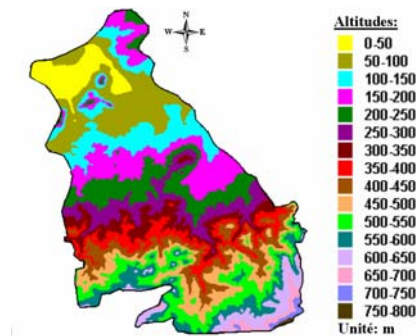


Fig. 3. MNA du bassin versant d'Oued el Maleh.

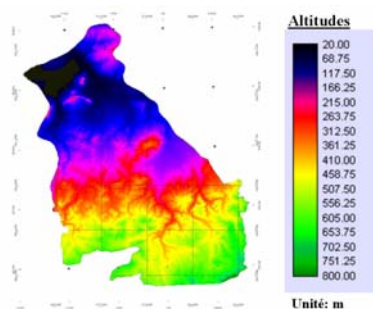


Fig. 4. MNT du bassin d'Oued el Maleh

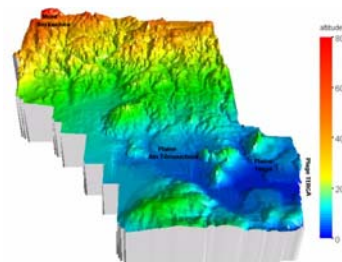


Fig. 5. MNT du bassin d'Oued El Maleh en 3Dimensions

### c) Indice de pente $I_p$

Cet indice se calcule à partir du rectangle équivalent. Il est égal à la somme des racines carrées des pentes moyennes de chacun des éléments pondérés par la surface intéressée, soit :

$$I_p = L^{-1/2} \cdot \sum_n \sqrt{\alpha_i \cdot d_i} \quad (2)$$

$L$ : Longueur du rectangle équivalent,  $\alpha_i$  présente la fraction en % de la surface comprise entre deux courbes de niveau voisines distantes de  $d_i$ , cette surface a été extraite automatiquement à partir des requêtes géographiques basées sur les données du modèle numérique d'altitude, cette méthode donne des résultats fiables.

Indice de pente de Roche  $I_p = 3.85$

Indice global de pente : Sur la courbe hypsométrique, on prend les points tels que la surface supérieure ou inférieure soit égale à 5% de  $A$ . On en déduit les altitudes  $H_5$  et  $H_{95}$  entre lesquelles s'inscrit 90% de l'aire du bassin et la dénivelée  $D = H_5 - H_{95}$

L'indice global est égal à:  $IG = 11.63$

$$I_G = \frac{D}{L} \quad (3)$$

Nous avons utilisé la dénivelée spécifique ( $D_s$ ) et ce afin de classer le relief du bassin d'étude. Sa valeur est Calculée par la formule suivante :  $D_s = IG \cdot \text{racine } A$  (Bouanani, A 1997), elle est égale à **326.22**. Le bassin versant d'oued El Maleh se distingue par un relief **fort**.

### 4.1.3 Les caractéristiques du réseau hydrographique

Le réseau hydrographique peut prendre une multitude de formes, la différenciation des réseaux est due aux facteurs géologiques, climatiques, topographiques et humains (Bouanani, A 1997).

Ce réseau est caractérisé par sa hiérarchisation et son développement. L'étude de ces deux éléments est très nécessaire pour évaluer l'écoulement des eaux superficielles, et peut se faire très aisément en utilisant le langage SQL.

### Hiérarchisation et mise à jour du réseau d'oued El Maleh

La méthode utilisée pour l'extraction du réseau hydrographique nous a amené à digitaliser les cours d'eau sur les cartes topographiques concernant la région à 1/25.000 (année 1986) d'une part, et de superposer ce fichier numérique sur des images satellitaires Landsat ETM+ prise en année 2002 d'autre part.

La superposition de ces deux couches d'informations, nous permet d'actualiser le réseau hydrographique du bassin versant d'oued El Maleh.

Le réseau actualisé sera hiérarchisé selon la méthode de STRAHLER (1957).

Ensuite nous avons introduit à l'entité (R\_Hydro) une base de données contenant des informations sur les attributs (code, nom\_BV, ordre), facilitant ainsi son traitement automatique.

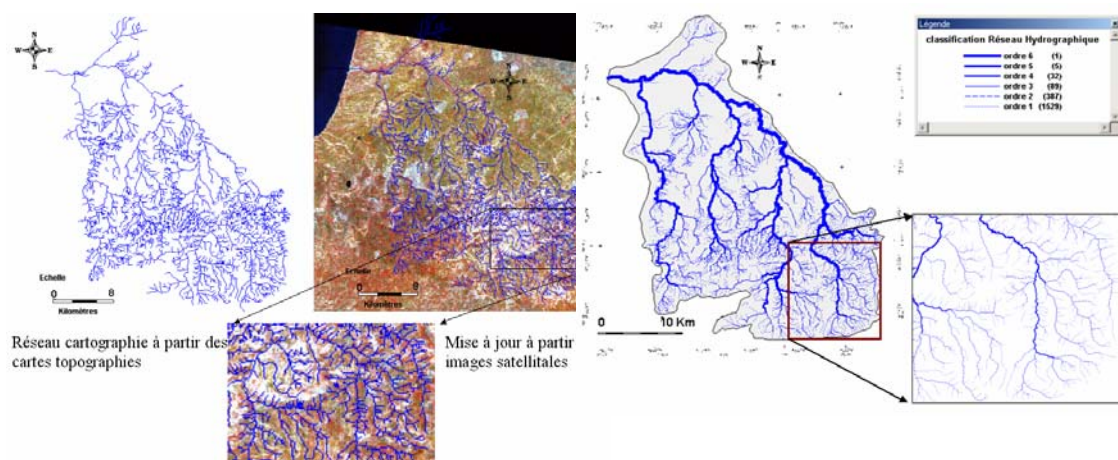


Fig. 6. Mise à jour du réseau hydrographique

Fig. 7. Classification du réseau hydrographique par la méthode Strahler (1957).

### Caractéristiques du réseau :

Le réseau hydrographique est caractérisé par la longueur du cours d'eau principal ( $L$ ), c'est la distance curviligne depuis l'exutoire jusqu'à la ligne de partage des eaux et cette dimension est obtenue tout simplement en faisant une requête géographique, la longueur donc sera égale à **65.1 km**.

Quant au profil longitudinal du cours d'eau principal, à l'aide des modèles numériques de terrain, sa représentation automatique est devenue aisée et donne des informations fiables.

Nous avons établi cette représentation en utilisant les systèmes d'information géographique et le modèle numérique d'altitude.

Le graphe obtenu démontre qu'on est en présence, à l'amont du bassin, d'une rupture de pente entre les altitudes 350 et 400 m, reflétant ainsi une structure particulière du sol, suivie vers l'aval d'une pente douce.

L'étude de développement d'un réseau hydrographique nous détermine l'influence de la topographie et la structure géologique sur son évolution, aux moyens d'indices calculés en utilisant le langage SQL. Ces indices sont : la densité de drainage, la densité hydrographique, le rapport de confluence, le rapport de longueur, la fréquence des cours d'eau et enfin le coefficient de torientialité.

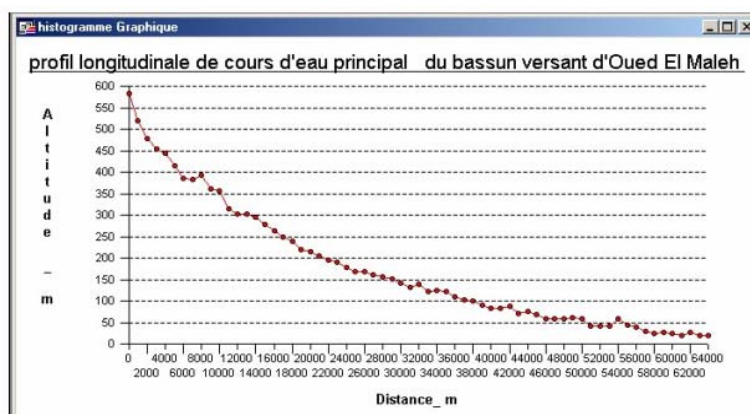


Fig. 8 profil en long d'un cours d'eau principal

#### 4.1.4 Le Degré de développement du réseau :

La densité de drainage au niveau d'oued El Maleh est égale à **1.79 km/km<sup>2</sup>**, ceci traduit l'existence de formations géologiques perméables, d'où un faible ruissellement et une forte infiltration.

La densité hydrographique est le nombre des cours d'eau par unité de surface avec une valeur de 2.6, nous pouvons conclure que le bassin est mal arrosé favorisant l'infiltration par rapport au ruissellement.

Pour ce qui est des rapports de longueur, nous nous sommes basés sur la classification des cours d'eau, selon Horton (1932) et Schumm (1956), nous avons introduit la **requête SQL** suivante : **Select Nom\_bv, Ordre, L\_tot\_km, Nombre, longueur\_km/nombre ``L\_moy\_km``from hiér\_O.el maleh by Ordre order by ordre into rapport\_de\_longueur** Browse\* from rapport\_de\_longueur

Table 3.Extraction automatique du rapport de longueur en utilisant le langage SQL

rapport_de_longueur Données						
	Nom_bv	Ordre	Nombre	L_tot_km	L_moy_km	RI
<input type="checkbox"/>	O.el maleh	1	1,573	3.40	0.43	1.70
<input type="checkbox"/>	O.el maleh	2	404	4.60	0.77	2.58
<input type="checkbox"/>	O.el maleh	3	94	7.60	1.99	1.32
<input type="checkbox"/>	O.el maleh	4	33	6.89	2.63	6.82
<input type="checkbox"/>	O.el maleh	5	5	9.74	17.94	2.74
<input type="checkbox"/>	O.el maleh	6	1	9.27	49.27	

Nous avons pris en considération, dans l'étude de développement, la fréquence du cours d'eau d'ordre 1 qui est estimée à **1.998**.

Cette valeur montre que le bassin versant présente une fréquence plus au moins faible, donc un réseau hydrographique d'une hiérarchisation moyenne.

Le dernier indice essentiel qui sert à caractériser le réseau hydrographique est le coefficient de torrentialité, c'est le rapport entre la fréquence des cours d'eau d'ordre 1 avec la densité de drainage et qui est de **3.576 Km/km<sup>4</sup>**.

Enfin, ces valeurs calculées par la méthode des requêtes géographiques donne à ce bassin le constat d'une hiérarchisation moyenne et un écoulement faible, ainsi qu'il est constitué par des formations perméables donc une infiltration importante.

## **Conclusion :**

Les caractéristiques physiques d'un bassin versant dépendent des différents facteurs climatiques, géologiques, topographiques et anthropiques.

L'étude de ces caractéristiques a été toujours établie par des méthodes classiques, nous avons montré, à partir de cette application, l'utilité des nouvelles techniques telles que les SIG et le langage SQL qui sont capables de gérer une base de données géographiques ainsi que l'apport rigoureux des modèles numériques de terrain dans les études de relief des bassins.

L'utilisation de ces méthodes nous permet un gain considérable en temps et en coût.

Les paramètres obtenus démontrent que notre bassin présente une forme allongée, un relief fort, un réseau hydrographique avec une hiérarchisation moyenne et un sol en général perméable ainsi qu'un temps de concentration de ruissellement faible.

## **References**

1. Boussema, M-R : intégration des données multisources pour l'étude du système hydrographique du delta de la Medjerda et de la Sebkhet er Riana télédétection et sécheresse , Aupelf-Uref John Libbey Eurotext Paris France (1990)
2. Bouanani, A : Hydrologie, transport solide et modélisation (étude de quelques sous bassin de la Tafna, NW – Algérie) Thèse de Doctorat, université Abou bekr belkaid Tlemcen, Algérie (1997)
3. Chaponniere, A : Fonctionnement hydrologique d'un bassin versant montagneux semi- aride cas du bassin versant du Rehraya (Haut Atlas marocain). Thèse de Doctorat, Centre d'étude spatiale de la Biosphère Paris–Grinon France (2005)
4. Dif, A: Modèle numérique de terrain et hydrologie : extraction des paramètres physiologiques et modélisation des écoulements superficiels d'un bassin versant à partir d'un MNT. Application à l'oued El Besbes (BV de la grande sebkha d'Oran) Thèse de magistère Centre National des Techniques Spatiales d'Arzew, Oran Algérie (1999)
5. Dridi, B : Impact des aménagements sur la disponibilité des eaux de surface dans le bassin versant du Merguellil (Tunisie centrale). Thèse de Doctorat, université de louis Pasteur centre d'étude et de recherche eco-géographique Strasbourg (2000)
6. Nouvelot, J-F : Guide des pratiques hydrologiques sur les petits bassins versants ruraux en Afrique tropical et équatoriale. ORSTOM Montpellier France (1989)
7. Le pape, S : analyse et quantification du réseau hydrographique, le réseau hydrographique comme objet vectoriel. Mémoire d'Ingénieur, Ecole supérieure des géomètres et topographies le Mans France(1998)

## **A Tool for Semantic Integration of Geographic Ontologies**

Djerriri khelifa<sup>1</sup>, Malki Mimoun<sup>2</sup>

<sup>1</sup>Earth Observation Division, Center for Spatial Techniques, Arzew, Algeria  
Djerriri@yahoo.fr

<sup>2</sup>EEDIS Laboratory, Sidi Bel Abbès University, Sidi Bel Abbès, Algeria  
Malki\_m@yahoo.fr

**Abstract.** Reusing data produced by other sources is the main motivation of System interoperability and data integration; enterprises tend to reduce their investment for producing data by integrating outsourced data.

The present work aims at resolving semantic conflicts during Geo-Spatial databases schema integration and reducing the cost of generating global schemata for heterogeneous local databases. It focuses on the semantics related to the terms used as identifiers in schema definitions.

We consider ontologies as knowledge structures that specify terms, their properties and relations among them to enable integration. The presented approach integrates schemata of different geographic information systems (GIS), where each GIS is using its own Geo-ontology. To reach our goal the concept of Semantic similarity is used and the Similarity relations between heterogeneous ontologies are discovered by reasoning system using the high level ontology WordNet. The output of the system is semantic similarity measure that is a confidence score that reflects the semantic relation between the meanings of two definitions.

**Keywords:** interoperability, Data integration, GIS, WordNet

### **1 Introduction**

Today we see a high proliferation of geographic information systems (GIS) and large amount of geographic data being collected and stored by different organizations. But, very often the geographic information required by a data consumer may not be available from a single data source and needs to be retrieved from multiple data sources. Reusing data produced by other sources is the main motivation of System interoperability and data integration; enterprises tend to reduce their investment for producing data by integrating outsourced data. However, integration as precondition of data reuse demands its own costs. Data integration has to resolve the differences in data structures, as well as, solving semantic heterogeneities.

The present work contributes to handling semantic heterogeneities during GIS databases schema integration. The approach aims at reducing the cost of generating global schemata for GIS databases. The focus of the work is on the semantics related to the terms used as identifiers in schema definitions.

The semantics of an information source may be described using an ontology defined as “an explicit specification of conceptualization”. The task of integration using different ontologies is a classical problem in information science, and continues to be highly active research issue within many topics, including databases, interoperability, the semantic Web, knowledge representation, data warehousing, and, of course geographical information integration.

Our work shows how similarity relations are discovered by reasoning system using high level ontologies or dictionaries. To reach our goals we used the concept of Semantic similarity which is defined as the confidence score that reflects the semantic relation between the meanings of two definitions, even it is difficult to gain a high accuracy score because the exact semantic meanings are completely understood only in a particular context.

One of the challenges is to define the mappings in order to integrate the different sources. We explore some alignment techniques to generate semantic mappings automatically. A tool, developed using CSharp programming language and the WordNet.Net framework, finds mappings or suggests indicators in order to help users find mappings. Experiment results on urban geographical information systems domain are discussed.



The content of this article is structured as follow: section 2 gives motivation of our research. Section 3 is a review of existing research related to semantic interoperability. Section 4 presents our approach and experiments. Section 5 concludes this article and discusses future works.

## **2 Motivation**

Semantic heterogeneity between geospatial data from different sources constitutes one of the major problems to the integration of these data. Geo-spatial database models are similar to the observation made by human beings. Two persons observing the same geographic area or the same satellite image may form two completely different models describing these data. Similarly, two geo-database schemas S1 and S2 may be quite different, even though both of them are inherently representing the same underlying data. It might even be difficult to say which of S1 or S2 is better. This relative usefulness would depend on the particular application for which the model is used. For example Highways in one model could be named as Expressways in the second model the same thing can occur between the two concepts Forest and wooded area. On one hand, proposing a standard model to be adopted by all data producers would lead to loss in the resolution of the collected data. A more feasible approach is to allow for the different data producers to maintain their own model locally while specifying how concepts in their own model correspond to concepts in the others models or in a global model. On the other hand, we note that even when data producers use a particular standard, their exact implementation of that standard will differ.

One of the solutions to the Semantic heterogeneity problem consists in developing ontologies to describe semantic of the geospatial data. Ontology forms a description of an abstract model and of the concepts which are used. However, ontologies are also heterogeneous, since they often differ according to their level of abstraction, their terminology, their structure, the definition of concepts, etc. In this case, the mapping or semantic integration of ontologies is a necessary condition to semantic interoperability. The integration of ontologies is the process of forming an ontology for a given subject by the re-use of several ontologies describing different subjects. The integration of ontologies can be carried out by the mapping, the alignment or the fusion of ontologies, these processes representing increasing degrees of integration of ontologies. In this article, we focus explicitly on mapping approaches.

The study of ontologies began in the domain of philosophy but found extensive application in many diverse disciplines such as artificial intelligence and Information technology. In the past few years, the subject of ontology has become one of the most important fields of research in geographical information science and many researchers have been focused on developing ontologies for geographic features.

According to Gruber's definition an ontology is «a formal, explicit specification of shared conceptualization» [1]. Studer et al. provide an explanation of this definition: 'conceptualization' is an abstract model of some phenomena in the world usually restricted to a particular domain; 'explicit' means that the type of concepts used and the constraints on their use are given explicit definitions; 'formal' refers to the fact that the ontology should be machine-readable; and 'shared' means that an ontology is agreed between a group of the domain experts.

The primary components of an ontology are 'concepts', 'axioms', and 'relations'. Concepts refer to a set of objects in a domain. Relations represent types of interaction among the concepts in an ontology and axioms refer to constraints imposed on concepts or on relations. Therefore, ontology is the study of explaining reality by representing it as concepts, relations and axioms.

An ontology ranges from a simple taxonomy to an axiomatized set of concepts and relationship types. In the current work, we focus on taxonomies for land use in urban areas.

There are a number of land cover standards which can be considered as simple geospatial ontologies:

AKTIS (Amtliches Topographisch-Kartographisches Informations System) is an official information system in Germany. The AKTIS standard focuses more on different types of areas used for human activities such as those used for transportation and vegetated areas.

CORINE (Co-ordination of Information on Environment) standard presents a more detailed hierarchical land cover classification compared to the AKTIS standard. For instance the CORINE taxonomy contains many different kinds of land-use areas which are not further differentiated in the AKTIS catalogue. Therefore, the land cover ontologies implemented in this article is based on the CORINE and AKTIS classifications.



### **3 Ontology alignment and mapping**

The mapping of ontologies consists in identifying a formal expression which describes the semantic relation between two concepts belonging to different ontologies. Consequently, the mapping of ontologies is closely related to the concept of semantic similarity. The majority of mapping methods use a mapping model which is based on a semantic similarity model to identify the semantic relations between the entities of two ontologies. The mapping model is thus at the heart of the process of integration.

The different types of matching that can between two heterogeneous ontologies occur are as follows:

**Naming:** Such instances can be seen when the name of same concepts, proprieties or even relations, is different in two ontologies like Forest and wooded area. In the reverse case it could also be that different entities (concepts proprieties or relations) are named the same in the two ontologies, this later case happens too rarely in geo-ontologies.

**Missing:** Some entities occurring in some ontology could be obsolete or unknown for another ontologies.

**Evolving:** As a reverse of above type, there could be entities which are newly introduced in a ontology and are not a part of other ontologies.

**Complex:** In this case, entities occurring in one ontology could be a mix of two or more entities in other ontologies. One could also identify the reverse case where the entities are only a part of the complex concepts, proprieties or relations of other ontologies. (Eg. One-Way-Roads is a only a part of the complex concept of Roads which include all other roads)

### **4. Related Work**

Rodriguez, Egenhofer and Rugg proposed an approach for assessing similarities among geospatial feature class definitions. The similarity evaluation is basically done over the semantic interrelation among classes. In that sense, they consider not only the taxonomic relations and the part-Of relations but also distinguish features (parts, functions and attributes) [3].

Fonseca proposed an ontology-driven GIS architecture to enable geographic information integration. In that proposal, the ontology acts as a model-independent system integrator. The work of Fonseca et al focuses on the application level, in which they can work on the translation of a conceptual schema to application ontology [4].

Hakimpour and Timpf proposed the use of ontology in the resolution of semantic heterogeneities especially those found in Geographic Information Systems. The goal was to establish equivalences between conceptual schemas or local ontologies. Basically the process is done in two phases: First, a reasoning system is used to merge formal ontologies. In the last phase, they find the possible meaningful mappings in the generated global schema and by that establish the mapping of data between the databases. Then the data (instances) from the local schemas are mapped [5].

Stoimenov and Djordjevic-Kajan propose the GeoNis framework to reach the semantic GIS data interoperability. It is based on mediators, wrappers and ontologies. The use of ontologies was proposed as a knowledge base to solve semantic conflicts as homonyms, synonyms and taxonomic heterogeneities [6].

### **5. Our Experiments**

For testing our approach we chose Sidi Bel Abbes city as study area, this choice is due essentially to diversity of themes present in this area, which it contains building, water streams, vegetated areas, and roads. It considers then as an ideal area for testing our approach.

We have as data describing our study area a high resolution satellite image, a scanned plan, attributes data and of course our knowledge of the city.

The following figures represent the satellite images of the city center and its scanned plan.



Fig. 2. Satellite Image and scanned map of the study area

Using the previous available data we tried to build two Geographic information systems modeling the same geographic area. The result GIS are based on two different data models and using two heterogeneous land-cover classifications systems which are shown on the followings figures:

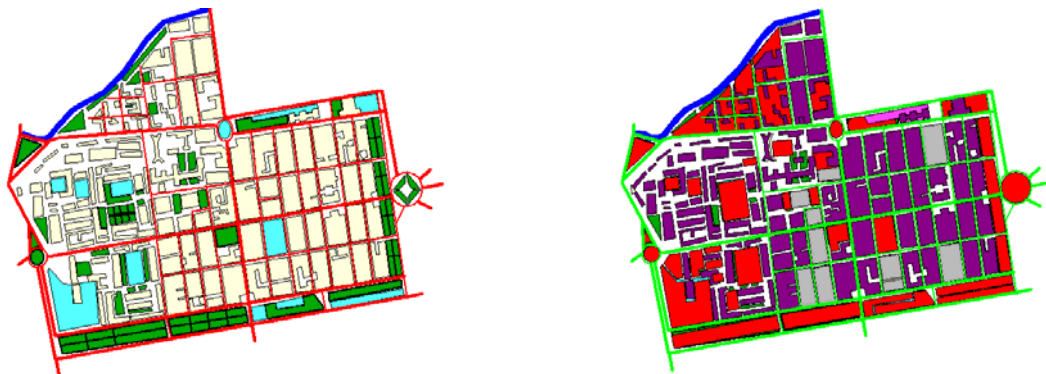


Fig. 2. Graphical representation of the result GIS

The first GIS built for our area contains the followings information layers: Built Space layer, non-Built Space layer, Water Stream layer, Transportation Infrastructure layer.

While the second GIS contains: Building Area layer, non-building area layer, Stream of water layer, Road network layer, Green Area layer.

We can see clearly that the both GISs are semantically heterogeneous and use two different nomenclatures. In fact it is easy to construct simple ontologies corresponding to them by following the steps bellow:

Step 1 : ontology domain definition

Step 3 : enumerate the most important terms in the domain

Step 4 : define the classes and their hierarchies

Step 5 : definie the classes attributes

Step 6 : define the attributs proprieties

- attributes cardinalities
- attributes value type
- domain and rang of attributes

Step 7 : create instances

By applying these steps, we obtained the two ontology hierarchies below:

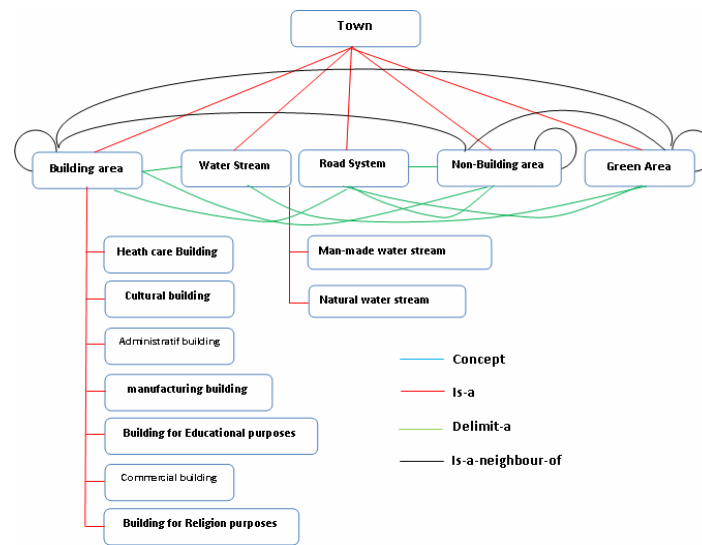


Fig. 3. First Ontology hierarchy

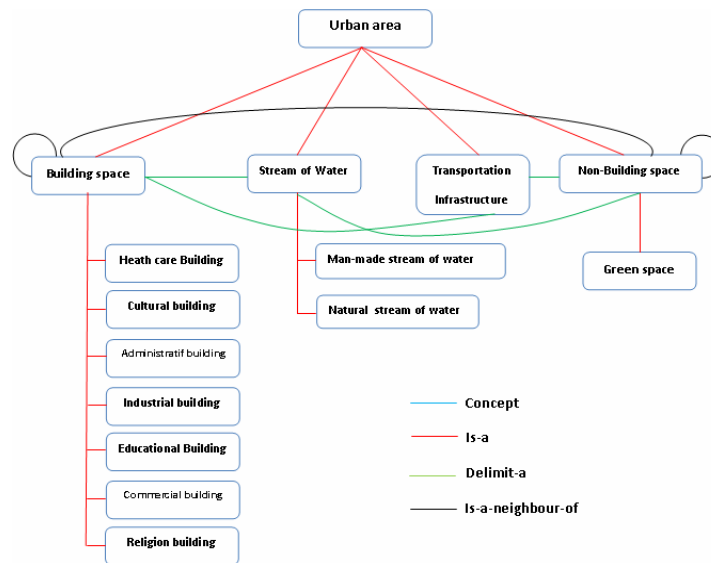


Fig. 4. Second Ontology hierarchy

## 5.2 Syntactic Similarity

We conduct our first evaluation using syntactic similarity measurements. The similarity is calculated in three steps:

- Partition each concept name into a list of tokens: A concept name is a list of words or abbreviations; it may be composed without separator characters. Take an example: 'building area' is being tokenized as "area" and "building".
- Computing the similarity between tokens by using a string edit-distance algorithm : the method uses an edit-distance string matching algorithm: Levenshtein distance is obtained by finding the

cheapest way to transform one string into another. Transformations are the one-step operations of insertion, deletion and substitution. In the simplest version substitutions cost about two units except when the source and target are identical, in which case the cost is zero

- Computing the similarity between two token lists: After splitting each concept into token lists, we capture the similarity between two concept names by computing the similarity of those two token lists, which is reduced to the *bipartite graph matching* problem.

The problem can now be described as follows : Given a graph  $G(V,E)$ ,  $G$  can be partitioned into two sets of disjoint nodes  $X$ (left) and  $Y$  (right) such that every edge connects a node in  $X$  with a node in  $Y$ , and each edge has a non-negative weight.

- $X$  is the set of the first list of tokens.
- $Y$  is the set of the second list of tokens.

$E$  is a set of edges connecting between each couple of vertex  $(X,Y)$ , the weight of each edge which connects an  $x_1$  to a  $y_1$  is computed by the similarity of  $x_1$  token and  $y_1$  token.

The task is to find a subset of node-disjoint edges that has the *maximum total weight*. The similarity of two strings is computed by the number of matched strings.

The result of syntactic similarity between the concepts names of our taxonomies is shown on the table below

**Table 1.** Syntactic similarity scores obtained in the second experiment.

	Town	Building area	Non-building area	Water stream	Road system	Green area	Health care building	Cultural building	Administratif building	Industrial building	Educational building	Commercial building	Building for worshipping	Artificial water stream	Natural water stream
Urban area	0,11	0,50	0,38	0,36	0,30	0,67	0,22	0,19	0,19	0,22	0,21	0,22	0,14	0,24	0,28
Building space	0,08	0,69	0,56	0,08	0,23	0,15	0,56	0,50	0,48	0,50	0,47	0,50	0,48	0,14	0,06
Health care building	0,06	0,56	0,56	0,17	0,11	0,17	1,00	0,56	0,43	0,50	0,53	0,56	0,48	0,24	0,17
Cultural building	0,06	0,56	0,56	0,13	0,13	0,13	0,56	1,00	0,48	0,67	0,63	0,67	0,43	0,19	0,28
Administratif building	0,05	0,48	0,48	0,14	0,14	0,14	0,43	0,48	1,00	0,57	0,48	0,48	0,43	0,24	0,14
Industrial building	0,06	0,56	0,56	0,22	0,17	0,17	0,50	0,67	0,57	1,00	0,63	0,61	0,43	0,29	0,22
Educational building	0,11	0,53	0,53	0,11	0,11	0,16	0,53	0,63	0,48	0,63	1,00	0,53	0,43	0,19	0,21
Commercial building	0,06	0,56	0,56	0,11	0,11	0,17	0,56	0,67	0,48	0,61	0,53	1,00	0,48	0,19	0,17
Religion building	0,06	0,56	0,56	0,00	0,06	0,13	0,50	0,56	0,48	0,50	0,58	0,50	0,52	0,14	0,00
Non-built space	0,07	0,14	0,56	0,07	0,21	0,07	0,17	0,06	0,14	0,11	0,11	0,11	0,14	0,10	0,17
Stream of water	0,10	0,17	0,13	0,27	0,30	0,60	0,17	0,06	0,14	0,11	0,11	0,11	0,14	0,14	0,17
Green space	0,08	0,23	0,25	0,85	0,31	0,31	0,17	0,13	0,14	0,22	0,11	0,11	0,10	0,57	0,61
Man-made stream of water	0,05	0,14	0,29	0,52	0,19	0,19	0,14	0,10	0,14	0,19	0,10	0,10	0,10	0,57	0,62
Natural stream of water	0,05	0,15	0,25	0,55	0,20	0,20	0,15	0,25	0,14	0,20	0,20	0,15	0,10	0,67	0,90
Transportation infrastructure	0,11	0,11	0,14	0,21	0,18	0,14	0,18	0,14	0,14	0,21	0,11	0,11	0,14	0,21	0,25

We can see clearly that the syntactic matcher failed in finding the concepts that have the same meaning and different syntactic forms like urban area and town.

## 5.2 Semantic Similarity

In the previous experiment, we have presented an approach for capturing similarity between concepts of two ontologies that was concerned with the syntactic similarity. In this second experiment we discuss another approach that is more concerned with the meaning of concepts. Semantic similarity is a

confidence score that reflects the semantic relation between the meanings of two definitions. It is difficult to gain a high accuracy score because the exact semantic meanings are completely understood only in a particular context.

The goal of this experiment is to present a dictionary-based algorithm to capture the semantic similarity between concepts, which is heavily based on the WordNet semantic dictionary.

WordNet is a lexical database which is available online and provides a large repository of English lexical items. There is a multilingual WordNet for European languages which are structured in the same way as the English language WordNet [7]. WordNet was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. Synsets are equivalent to senses = structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept it represents. Synsets are connected to one another through the explicit semantic relations. Some of these relations (hypernym, hyponym for nouns and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies. Our implementation of the WordNet based semantic similarity system is based on WordNet.Net which is an open-source .NET Framework library.

Given two concepts, the measurement determines how similar the meaning of two concepts is. The higher the score the more similar the meaning of the two concepts.

The steps for computing semantic similarity are the followings :

- First each concept is partitioned into a list of tokens.
- Part-of-speech disambiguation or tagging).
- Stemming words.
- Find the most appropriate sense for every word in a concept (Word Sense Disambiguation).
- Finally, compute the similarity of the concepts based on the similarity of the pairs of words.

**Tagging part of speech** is to identify the correct part of speech (POS - like noun, verb, pronoun, adverb ...) of each word in the whole concept. The algorithm takes a concept as input and a specified a finite list of POS tags. The output is a single best POS tag for each word.

**Stemming** is a process of removing the common morphological and inflexional endings of words. It can be thought of as a lexicon finite state transducer with the following steps: Surface form -> split word into possible morphemes -> getting intermediate form -> map stems to categories and affixes to meaning -> underlying form. For example: areas -> area + s -> areas.

**Semantic relatedness and Word Sense Disambiguation:** To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words.

**Semantic similarity between two synsets:** In WordNet, each part of speech words (nouns/verbs...) are organized into taxonomies where each node is a set of synonyms (synset) represented in one sense. If a word has more than one sense, it will appear in multiple synsets at various locations in the taxonomy.

A simple way to measure the semantic similarity between two synsets is to treat taxonomy as an undirected graph and measure the distance between them in WordNet. The shorter the path from one node to another, the more similar they are. The length of the path between two members of the same synsets is 1 (synonym relations).

There are many proposals for measuring semantic similarity between two synsets: Wu & Palmer, Leacock & Chodorow, P. Resnik. In this work, we experimented simple measurements.

We will now describe the overall strategy to capture semantic similarity between two concepts. Given two concept names  $C1$  and  $C2$ , we denote  $m$  to be length of  $C1$ ,  $n$  to be length of  $C2$ . The major steps can be described as follows:

1. Tokenization.
2. Perform word stemming.
3. Perform part of speech tagging.
4. Word sense disambiguation.
5. Building a semantic similarity relative matrix  $R[m, n]$  of each pair of word senses, where  $R[i, j]$  is the semantic similarity between the most appropriate sense of word at position  $i$  of  $C1$  and the most appropriate sense of word at position  $j$  of  $C2$ . Thus,  $R[i, j]$  is also the weight of edge connect from  $i$  to  $j$ . If a word does not exist in the dictionary we use the edit-distance similarity instead and output a lower associated weight.

We formulate the problem of capturing semantic similarity between concepts as the problem of computing a maximum total matching weight of a bipartite graph, where  $C1$  and  $C2$  are two sets of disjoint nodes. We use the Hungarian method to solve this problem.

6. The match results from the previous step are combined into a single similarity value for two concepts. There are many strategies to acquire an overall combined similarity value for sets of matching pairs. Matching average: where  $match(C1, C2)$  are the matching word tokens between  $C1$  and  $C2$ . This similarity is computed by dividing the sum of similarity values of all match candidates of both concepts  $C1$  and  $C2$  by the total number of set tokens.

The result of syntactic similarity between the concepts names of our taxonomies is shown on the table below.

**Table 2.** Semantic similarity scores obtained in the second experiment.

	Town	Building area	Non-building area	Water stream	Road system	Green area	Health care building	Cultural building	Administratif building	Industrial building	Educational building	Commercial building	Building for worshipping	Artificial water stream	Natural water stream
Urban area	0,54	0,82	0,69	0,57	0,69	0,78	0,65	0,56	0,53	0,58	0,54	0,59	0,54	0,51	0,54
Building space	0,74	0,96	0,79	0,69	0,70	0,84	0,82	0,70	0,71	0,72	0,74	0,78	0,74	0,59	0,68
Health care building	0,59	0,85	0,73	0,69	0,67	0,70	1,00	0,68	0,65	0,67	0,69	0,71	0,69	0,61	0,70
Cultural building	0,57	0,75	0,62	0,57	0,56	0,56	0,68	1,00	0,58	0,70	0,68	0,70	0,61	0,49	0,66
Administratif building	0,56	0,75	0,63	0,56	0,57	0,57	0,65	0,58	1,00	0,66	0,61	0,64	0,60	0,53	0,58
Industrial building	0,57	0,77	0,64	0,61	0,58	0,58	0,67	0,70	0,66	1,00	0,68	0,68	0,63	0,54	0,61
Educational building	0,59	0,78	0,66	0,55	0,58	0,58	0,69	0,68	0,61	0,68	1,00	0,67	0,64	0,51	0,60
Commercial building	0,66	0,82	0,68	0,61	0,65	0,63	0,71	0,70	0,64	0,68	0,67	1,00	0,68	0,56	0,63
Religion building	0,74	0,90	0,77	0,61	0,69	0,72	0,77	0,70	0,71	0,72	0,74	0,77	0,74	0,54	0,61
Non-built space	0,44	0,79	0,97	0,47	0,44	0,59	0,67	0,61	0,59	0,60	0,61	0,65	0,61	0,41	0,49
Stream of water	0,68	0,81	0,68	0,77	0,61	0,96	0,62	0,44	0,47	0,46	0,46	0,55	0,49	0,63	0,74
Green space	0,42	0,69	0,55	1,00	0,68	0,77	0,69	0,57	0,56	0,61	0,55	0,61	0,53	0,84	0,89
Man-made stream of water	0,41	0,79	0,72	0,85	0,65	0,69	0,78	0,71	0,70	0,73	0,69	0,74	0,78	0,75	0,86
Natural stream of water	0,41	0,68	0,59	0,89	0,63	0,74	0,70	0,66	0,58	0,61	0,60	0,63	0,55	0,79	1,00
Transportation infrastructure	0,39	0,78	0,81	0,76	0,65	0,66	0,71	0,60	0,62	0,64	0,60	0,65	0,59	0,64	0,72

## **6. Conclusion and future work**

The challenge faced here was to develop a methodology that achieves good practical results when integrating two geographic ontologies by measuring their content similarities. To achieve this goal we have tested in this work two different kinds of similarity measurements between ontologies, the first was based only on the syntactic forms of the taxonomies concept names, while the second was based on the using of external high level ontology.

Both experiments are developed using CSharp programming language and the WordNet.Net framework library. By analyzing the obtained results we believe that the alignment of geospatial ontologies based on the semantic similarity is more realistic and much closer to human being reasoning.

As future work, we plan to study the impact of the features attributes, taxonomy, conventional and topological relationships and spatial relationships, such as distance relations, on the similarity measure between two ontologies.

## **References**

1. Gruber T., Towards Principles for the Design of Ontologies Used for Knowledge Sharing, *International Journal of Human and Computer Studies*, 43(5/6), p. 907-928, 1995
2. Studer R., V.R. Benjamins, D. Fensel. *Knowledge Engineering: Principles and Methods*. Data and Knowledge Engineering, vol.25. pp. 161-197, 1998
3. Rodriguez, M. A. and Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Trans. Knowl. Data Eng.*, 15(2):442– 456.
4. Fonseca, F., Egenhofer, M., Agouris, P., and Camara, G. (2002). Using ontologies for integrated geographic information systems. *Transactions in Geographic Information Systems*, 6(3).
5. Hakimpour, F. and Geppert, A. (2002). Global schema generation using formal ontologies. In Spaccapietra, S., March, S. T., and Kambayashi, Y., editors, *ER*, volume 2503 of *Lecture Notes in Computer Science*, pages 307–321. Springer.
6. Stoimenov, L. and Djordjevic-Kajan, S. (2005). An architecture for interoperable GIS use in a local community environment. *Computers and Geosciences*, 31:211–220.
7. Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(1 1):39– 41.

*Natural language processing*

---





## Classification des Textes Arabes Basée sur l'Algorithme des Arbres de Décision

Fouzi Harrag <sup>1</sup>, Aboubaker Hamdi-Chrif <sup>2</sup>

<sup>1</sup>Département d'informatique  
Université Farhat Abbas, Sétif – Algérie  
[hfouzi2001@yahoo.fr](mailto:hfouzi2001@yahoo.fr)

<sup>2</sup> Computer College, Qassim university  
Buraydah-Saudi Arabia  
[elhamdi62@gmail.com](mailto:elhamdi62@gmail.com)

**Résumé:** Cet article présente les résultats de l'application de l'algorithme des arbres de décision pour la classification des textes arabes. Les Expérimentations sont déroulées sur deux corpus en langue arabe et les résultats montrent que l'utilisation hybride du seuil de la fréquence des termes *TFT (Term Frequency Thresholding)* avec le critère du gain de l'information *IG (Information Gain)*, déjà existant dans l'algorithme des arbres de décision lui-même, forme l'approche préférable pour la sélection des attributs. L'étude conclue que l'amélioration de notre classificateur arbre de décision donne de très bons résultats du point de vue efficacité et performance. L'exactitude générale est autours du 0.93 pour le corpus scientifique et du 0.91 pour le corpus littéraire. Nous concluons aussi que l'augmentation de la taille de l'ensemble d'apprentissage et la nature du Corpus ont une grande influence sur l'amélioration des performances du classificateur. Ainsi la réduction des erreurs de classification nous oblige de choisir soigneusement les documents de notre Corpus qui doivent appartenir à des catégories les plus thématiquement divergentes.

**Mots-clefs :** Algorithme arbre de décision, classification des textes, Corpus arabes, sélection d'attributs, Fouille de textes.

### 1 Introduction

Jour après jour, la masse d'information qui nous est disponible augmente. Internet s'enrichit continuellement de nouveau contenu, les entreprises emmagasinent de plus en plus dans leurs bases de données. Mais toute cette information serait sans intérêt si notre capacité à y accéder efficacement n'augmentait pas elle aussi. Pour en profiter au maximum, on a besoin d'outils et de techniques nous permettant de chercher, classer, conserver, mettre à jour et analyser ces données accessibles. La classification automatique de textes est l'une des solutions proposées pour ces problèmes, c'est un apprentissage qui vise à affecter une étiquette à un document en fonction de son contenu textuel. Les deux approches de la classification thématique sont l'approche supervisée et l'approche non supervisée. La classification thématique non supervisée consiste à regrouper les différentes données dans des regroupements les plus homogènes possibles [1]. Les algorithmes les plus connus de cette approche sont l'algorithme de Single-Pass [2], l'algorithme des K-moyennes [3], l'algorithme des nuées dynamiques [4] et Les méthodes neuronales [5] [6]. Les méthodes supervisées impliquent une intervention directe de l'utilisateur dans la constitution de l'ensemble d'apprentissage, l'algorithme de catégorisation projette l'information apprise dans la phase d'apprentissage sur un ensemble de test [7]. Parmi les algorithmes les plus connus de cette approche on retrouve L'algorithme des k les plus proches voisins [8], Le Classificateur bayésien naïf [9], Les Machines à support vectoriel (SVM) [10]. L'objectif de cet article est de présenter la démarche d'implémentation et d'évaluation d'un système de classification thématique supervisée des textes en langue arabe basée sur l'algorithme des arbres de décision. L'utilisation de deux corpus différents va permettre d'évaluer l'influence de la nature du corpus sur les performances du classificateur.

## 2 Classification Automatique des Textes

La première étape d'un système de classification est la préparation d'un ensemble de documents que l'on appelle corpus. Le prétraitement qui représente la seconde étape peut être subdivisé en plusieurs opérations. Initialement on procède à la reconnaissance des termes utilisés [11]. L'opération suivante est l'élimination des mots vides, elle peut se faire par une liste prédéfinie des mots non utiles (Stop List) ou par l'utilisation d'un étiqueteur morphosyntaxique. La troisième opération est l'application du processus de lemmatisation sur l'ensemble des textes de notre corpus. La lemmatisation permet de réduire les formes fléchies d'un mot en leur lemme. La deuxième étape ou indexation consiste à établir un lien pondéré pour la présence de chaque terme dans chaque document. Les documents résultants de cette opération sont divisés en deux ensembles: apprentissage et test. Les documents de l'ensemble d'apprentissage sont utilisés pour apprendre le système à reconnaître les différents modèles de catégories, tandis que les documents de l'ensemble de test sont utilisés pour évaluer le système, le processus de classification dépend de l'algorithme utilisé.

## 3 Classification par les Arbres de Décision

Les arbres de décision sont des techniques d'apprentissage supervisé. Les deux algorithmes les plus connus et utilisés sont CART [13] et C5 [14]. Ces algorithmes sont performants et génèrent des procédures de classification exprimables sous forme de règles. Pour classer un enregistrement par un arbre de décision il suffit de descendre dans l'arbre selon les réponses aux différents tests pour l'enregistrement considéré. Les règles des systèmes construits sont exhaustives et mutuellement exclusives.

### 3.1 Algorithme Arbres de Décision

Le schéma général de l'algorithme est donné :

**Donnée :** échantillon  $S$  de  $m$  enregistrements classés  $(x, c(x))$

**Initialisation :**  $A :=$  arbre vide ;

**Répéter** 1- Décider si le nœud courant est terminal.

**Si (terminal)** 2- Étiqueter le nœud courant par une feuille.

**Sinon (non terminal)** 3- Sélectionner un test.

**Jusqu'à (obtenir l'arbre de décision)** 4- Élaguer l'arbre de décision  $t$  obtenu.

**Sortie:** l'arbre  $t$  élagué.

L'idée centrale est de diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous ensembles d'exemples ne contenant (presque) que des exemples appartenant tous à une même classe. Les algorithmes des arbres de décision se différencient par les choix effectués pour le critère d'arrêt et pour le test de division (entropie ou gain). L'entropie est une valeur numérique quantifiant la notion de dispersion d'information (accroissement du désordre). Soit  $S$  un système partitionné en  $k$  sous systèmes alors son entropie  $E(S)$  est:

$$\text{Entropie}(S) = \text{Somme}_{(k=1..n)} p(k/p) * \log(p(k/p)) \quad (1)$$

Où :  $P(k/p) = N(k/p)/N(p)$  est la proportion d'éléments de classe  $k$  à la position  $p$ ,  $N(p)$  est le cardinal de l'ensemble des exemples associés au nœud  $p$  et  $N(k/p)$  : Est le cardinal de l'ensemble des exemples associés à  $p$  qui sont de classe  $k$ . Le gain est l'entropie du nœud actuel moins la somme pondérée de l'entropie des nœuds créés.

$$\text{Gain}(p, t) = i(p) - \text{Somme}_{(j=1..n)} p_j * i(p_j) \quad (2)$$

$i(p)$ : Représente l'entropie actuelle du nœud  $p$ . On souhaite obtenir l'entropie la plus faible car cela signifie qu'un grand nombre d'éléments appartiennent à la classe, on cherche donc à obtenir le gain maximum.

## 4 Méthodologie de Recherche

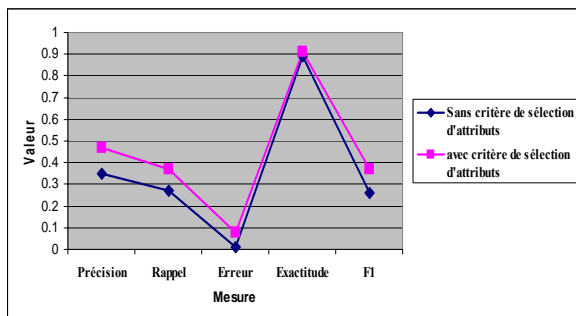
L'objectif de cet article est l'étude et l'évaluation du classificateur Arbre de Décision pour deux corpus de textes en langue arabe. Ces textes seront prétraités (nettoyage, élimination des mots vides et lemmatisation) puis soumis au processus d'indexation pour générer les ensembles d'apprentissage et de test. L'ensemble d'apprentissage sera utilisé pour la construction de l'arbre de décision qui sert à l'extraction des règles de classification. L'ensemble de test sera utilisé pour mesurer les performances de notre classificateur. La Précision, le Rappel, le taux d'Erreur, le taux d'Exactitude et la mesure F1 sont les mesures utilisées pour évaluer les résultats de la classification. Pour mieux évaluer notre classificateur nous avons opté pour l'utilisation de deux corpus différents, le "Corpus Littéraire" est un ensemble de traditions prophétiques (Hadiths du prophète 'qdlb') extraite à partir du CDROM de l'encyclopédie (Alkotob Altissâa الكتب التسعة) [15]. Il se caractérise par la spécialisation de son domaine 'Hadith'. Il comporte 453 documents répartis sur 14 catégories. Le "Corpus Scientifique" est un ensemble de textes arabes de différents domaines extraits du CDROM de l'encyclopédie scientifique arabe "Hal Tâalam هل تعلم؟" [16]. Il contient 373 documents répartis sur 8 catégories.

### 4.1 Expériences et Résultats

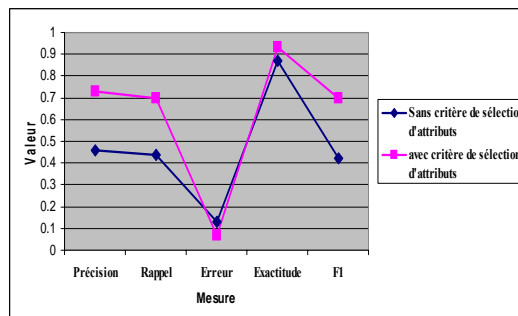
La première étape de ces expériences est l'extraction du lexique de l'ensemble des textes de deux corpus. Nous avons d'abord supprimé les mots vides en employons une liste de mots vides de la langue arabe. Sur le reste du lexique, nous avons appliqué un processus de lemmatisation totalement basé sur des règles de dépendances morphologiques.

#### 4.1.1 Expérience 1 : "Sélection d'Attributs"

Cette expérience consiste à évaluer l'influence du choix de critère de sélection d'attributs sur les performances du système de classification. Les lexiques extraits seront soumis à plusieurs opérations de filtrage et de sélection visant à diminuer substantiellement la taille. Nous avons testé les critères de la fréquence du terme ( $TF$ ), la fréquence du document ( $DF$ ) et la fréquence combinée ( $TF/DF$ ). Pour le corpus littéraire, la taille du vocabulaire retenue après l'application de ces critères est de 1938 termes. Pour le choix du meilleur seuil pour chacun de ces critères. Nous avons testés les seuils {2, 3} pour la fréquence  $TF$ , les seuils {14, 16} pour la fréquence  $DF$  et les seuils {(2/2), (2/4), (2/4), (2/5)} pour la fréquence combinée  $TF/DF$ . D'après les résultats, on constate que l'utilisation du critère  $TF$  pour un seuil égal à deux (2) donne le meilleur pourcentage 38 (%). Pour le corpus scientifique, la taille du vocabulaire retenue après l'application des critères de sélection d'attributs est de 1107 termes. D'après les résultats des tests réalisés, on constate que l'utilisation du critère  $TF$  pour un seuil égal à trois (3) donne le meilleur pourcentage 70(%). La comparaison des résultats de classification avec et sans utilisation du critère de sélection d'attributs pour les deux corpus est illustrée dans les figures 1 et 2. D'après la figure 1, on constate une amélioration de 12(%) pour la précision, de 10(%) pour le rappel, de 7(%) pour le taux d'erreur, de 2(%) pour le taux d'exactitude et de 11(%) pour la mesure F1. L'amélioration globale de performance pour le corpus littéraire est autour de 11(%)



**Fig. 1.** Influence des critères de sélection d'attributs sur les performances du classificateur pour un corpus littéraire de 340 documents

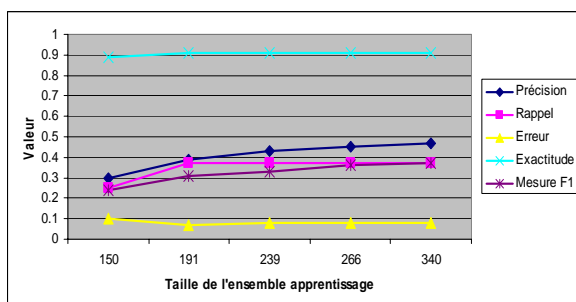


**Fig. 2.** Influence des critères de sélection d'attributs sur les performances du classificateur pour un corpus scientifique de 280 documents

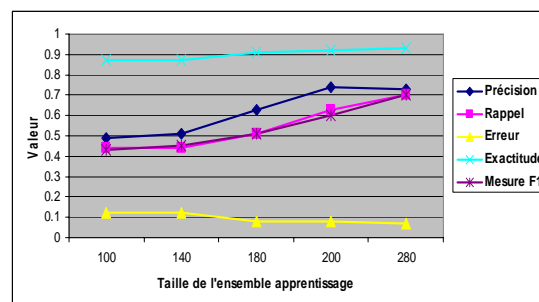
D'après la figure 2, on constate une amélioration de 27(%) pour la précision, de 26(%) pour le rappel, de 6(%) pour le taux d'erreur, de 6(%) pour le taux d'exactitude et de 28(%) pour la mesure F1. L'amélioration globale de performance pour le corpus scientifique est autour de 26.5 (%). Les résultats de cette expérience confirment que pour la phase de réduction de vocabulaire, la combinaison d'un critère de sélection d'attributs avec le critère du gain d'information, déjà existant dans l'algorithme des arbres de décision lui-même, permet d'améliorer considérablement les résultats de classification (16(%)).

#### 4.1.2 Expérience 2 : "Taille de l'Ensemble d'Apprentissage"

La deuxième expérience consiste à évaluer l'influence de la taille de l'ensemble d'apprentissage sur les performances du système de classification. Les différentes tailles testées sont: 150, 191, 239, 266, 340 pour le corpus littéraire et 100, 140, 180, 200, 280 pour le corpus scientifique. D'après les figures 3 et 4, on constate qu'il y a une amélioration globale de performance autour de 15(%) pour le corpus littéraire et de 26.5 (%) pour le corpus scientifique et il y a presque une stabilité pour les mesures d'erreur et d'exactitude. Les résultats de cette expérience confirment que l'augmentation de la taille de l'ensemble d'apprentissage a une grande Influence sur les performances du classificateur.



**Fig. 3.** Influence du critère de la taille d'ensemble d'apprentissage sur les performances du classificateur pour le corpus littéraire.



**Fig. 4.** Influence du critère de la taille d'ensemble d'apprentissage sur les performances du classificateur pour le corpus Scientifique.

#### 4.2 Comparaison des Résultats de l'Évaluation pour les Deux Corpus:

La comparaison des résultats d'évaluation montre que les meilleurs scores sont ceux des documents du corpus scientifique comme donnés par la figure 7.

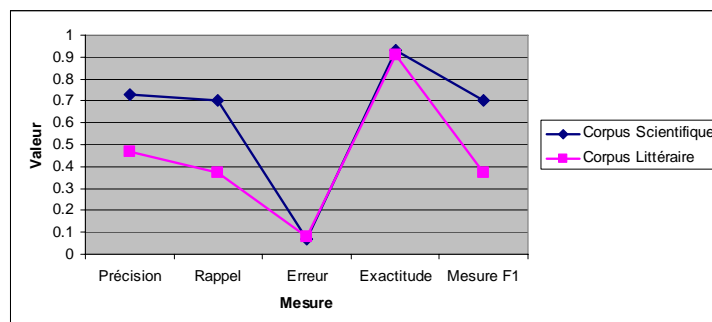


Fig. 5. Comparaison des résultats d'évaluation pour les deux corpus.

La faiblesse des résultats de l'évaluation du corpus littéraire nous a conduit à une révision des règles de classification qui a mis en évidence que la plupart des erreurs de classification sont en vérité dû à la nature et aux caractéristiques des documents du corpus littéraire. La plupart de ces documents contiennent un grand nombre de termes qui peuvent apparaître dans plus d'une catégorie ou d'une classe. D'autres facteurs peuvent aussi influencer sur les performances du système de classification telle que la taille des documents et la divergence thématique entre les différentes classes du corpus.

#### 4.3 Comparaison avec d'Autres Systèmes de Classification :

Les résultats de la comparaison de notre système avec d'autres systèmes de catégorisation des textes en langue arabe sont exposés dans la Table 1. La comparaison a été effectuée avec des systèmes probabilistiques basés sur l'algorithme bayésien naïf [18], statistiques basés sur l'algorithme du maximum [19] et linéaires basés sur le modèle de l'espace vectoriel et utilisant les mesures de di-similarité de *Dice*, de *Jaccard* et la mesure *Euclidienne* [20]. Cette comparaison montre que notre classificateur est l'un des meilleurs systèmes en terme de performance globale, il présente de meilleures valeurs pour les mesures de F1 et de précision ce qui prouve que notre système est plus précis. En terme de temps de classification, Les systèmes de classification basés sur les arbres de décision ont une bonne cote parmi les meilleurs systèmes. En effet le temps de classification de l'ensemble de test pour les deux corpus (93 documents pour le corpus scientifiques et 113 documents pour corpus littéraire) n'a pas dépasser les 120 secondes, très court par rapport au temps du système de classification basé sur l'algorithme K-PPV mentionné dans [21], car ce système a atteint un temps de 3004 seconds c à d environs 5,04 minutes.

Système	Précision	Rappel	Mesure F1
<b>Arbres de decision</b>	<b>73.00</b>	<b>70.00</b>	<b>70.00</b>
Bayésien naïf	67.88	71.96	67.83
Maximum entropie	50.00	84.20	62.7
Modèle de l'espace vectoriel (Dice)	41.00	44.00	42.00
Modèle de l'espace vectoriel (Jaccard)	54.00	61.00	57.00
Modèle de l'espace vectoriel (Euclidienne)	54.00	57.00	55.00

Table 1. Comparaison des résultats d'évaluation du classificateur des arbres de décision avec d'autres systèmes de classification.

## 5 Conclusion

Dans cet article, nous nous sommes intéressés à l'évaluation d'un système de classification basé sur l'algorithme des arbres de décision. Notre étude a été basée sur plusieurs expériences qui ont pour but d'étudier l'impact des critères de sélection d'attributs et de la taille de l'ensemble d'apprentissage sur les performances du système de classification présenté. L'utilisation de deux corpus différents nous a permis de conclure qu'une série de facteurs peut influencer sur le fonctionnement du système de classification en particulier la nature et la spécificité des documents de chaque corpus. Cette étude a aussi confirmé l'indépendance des méthodes de classification vis-à-vis de la langue utilisée. Pour nos futurs travaux, nous envisagerons de passer à la classification multi-classe par l'utilisation des arbres de décision floue et les techniques de segmentation thématique des textes. La dernière perspective est de mener d'autres expériences pour étudier l'impact de l'utilisation des extracteurs de racines, des analyseurs morphosyntaxiques, des dictionnaires et des thésaurus arabes sur les performances des systèmes de classification.

## Références

1. J. Callan. human language technologies, text categorization. (A power point presentation) <http://www.cs.waikato.ac.nz/>. 2004.
2. Forest, D. (2006) "Application de technique de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés", thèse de Doctorat en informatique cognitive, université du Québec à Montréal, Canada, 2006.
3. Rasmussen, (1992) E. Rasmussen, «Clustering Algorithms», in [Frakes & Bayes, « Information Retrieval, Data Structures & Algorithms », Prentice-Hall Inc., ISBN-0-13-463837-9. 1992], p. 419 à 442, 1992.
4. Mac Queen, (1967) J.B. Mac Queen, «Some methods for classification and analysis of multivariate observations», actes de 5th Berkeley Symposium on Mathematics Statistics and Probability, vol. 1, 1967.
5. Diday et al., (1982) E. Diday, J. Lemaire, J. Pouget, F. Testu, *Éléments d'Analyse des Données*, Dunod Informatique, 1982.
6. Kaski et al., (1998) S. Kaski, J. Kangas, T. Kohonen, «Bibliography of self-organized map (SOM) papers: 1981-1997», *Neural Computing Surveys*, vol. 1, n° 3 & 4, 1998, p. 1 à 176.
7. Kohonen, (2001) T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Science, Springer, 2001, ISBN 3-540-67921-9.
8. Simon, R. (2005) "Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés", mémoire de maîtrise en informatique, université de Laval, Québec, Canada, 2005.
9. Sebastiani.F, (1999) A Tutorial on Automated Text Categorisation. *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, pp. 7-35, 1999.
10. Yang. Y & Liu.X, (1999), A Re-examination of Text Categorization Methods. *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pp. 42-49, ACM Press, New York, US, 1999.
11. Vapnik. V.N. (1995), *The Nature of Statistical Learning Theory*. Springer, 1995.
12. Fox, C. (1992) " «Lexical analysis and stoplists», in in [Frakes & Bayes , « Information Retrieval, Data Structures & Algorithms », Prentice-Hall Inc., ISBN-0-13-463837-9. 1992], p. 102 à 130."
13. Salton G, (1983) & McGill. "Introduction to Modern Information Retrieval. Mc Graw-Hill, New York, 1983."
14. Breiman L. & Friedman J.H. & Olshen R.A & Stone, C.J. (1984) "Classification and regression trees. Technical report, Wadsworth International, Monterey, CA, 1984."
15. Quinlan J.R, C4.5, (1993) "Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993."
16. Salton G & Buckley C, (1988), "Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5), pp. 513-523, 1988."
17. The encyclopedia of the nine books for the honorable prophetic traditions, Sakhr company, 1997, <http://www.Harf.com>
18. El-Kourdi, M., Bensaid, A., Rachidi, T., Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics. August 28th. Geneva (2004).
19. Sawaf, H., Zaplo, J., Ney, H., Statistical Classification Methods for Arabic News Articles. *Arabic Natural Language Processing, Workshop on the ACL'2001*. Toulouse, France, July (2001).
20. Mohammed Naji Al-Kabi, Saja I. Al- Sinjilawi, A Comparative Study Of The Efficiency Of Different Measures To Classify Arabic Text, *University of Sharjah Journal of Pure & Applied Sciences* Volume 4, No. 2, June 2007.
21. M. M. Syiam, Z. T. Fayed & M. B. Habib, An Intelligent System For Arabic Text Categorization, *IJICIS*, Vol.6, No. 1, JANUARY 2006.

## Modèle p-contexte de classe pour la génération automatique des phrases arabe

Aouragh Si Lhoussain<sup>1</sup>, Allal Jelloul<sup>1</sup>, Yousfi Abdellah<sup>2</sup>

<sup>1</sup> Département de Mathématiques et informatique, Faculté des sciences,  
Université Mohamed premier, Oujda, Maroc

Mail : aouragh@hotmail.com, Allal@sciences.univ-oujda.ac.ma,

<sup>2</sup> Faculté des sciences juridique économique et sociale,  
Université Mohamed V, Souissi, Rabat, Maroc

Mail : yousfi240ma@yahoo.fr,

**Abstract.** Les modèles probabilistes de langage constituent les modèles de référence en traitement automatique de langage, ils permettent le calcul des probabilités de succession des mots à partir d'un corpus d'apprentissage. Ils peuvent être utilisés dans plusieurs domaines de TAL par exemple la traduction automatique, la reconnaissance de la parole continue, etc. L'utilisation des modèles p-contexte [Aouragh et al, 2006] pose un certain nombre de problèmes comme par exemple le nombre énorme de paramètres à estimer et le problème des événements non présentés dans le corpus d'apprentissage. Pour remédier à ces problèmes, nous proposons dans cet article une adaptation du modèle n-classe [Cerf-Danon et El-Bèze, 1991] à notre modèle p-contexte. L'utilisation de ce nouveau modèle a permis d'une part la réduction du nombre de paramètres à estimer et d'autre part la généralisation de certains événements non présentés dans le corpus d'apprentissage.

**Mots Clef:** modèle p-contexte, corpus d'apprentissage, syntaxe, sémantique, génération automatique des phrases.

### 1 Introduction

Les modèles de langage probabilistes utilisés jusqu'à présent ne résolvent pas de manière satisfaisante les nombreuses inflexions de la langue arabe.

Cet article décrit un modèle de langage statistique à base de classes sémantiques se basant sur les modèles p-contexte [Aouragh et al, 2005] pour la génération automatique des phrases arabes. Les modèles à base de classes ont au moins deux avantages : l'apprentissage du modèle est faite sur une bases de données réduites, et le nombre de paramètres à estimer est inférieur par rapport au modèle p-contexte.

### 2 Modèle p-contexte

La génération automatique des phrases est une opération difficile et délicate, elle nécessite une modélisation de tous les niveaux composant la phrase (syntaxe, sémantique,...). Le modèle p-contexte est un modèle proposé pour modéliser le niveau sémantique [Aouragh et al, 2005].

Soit  $w_{i_1} w_{i_2} \dots w_{i_n}$  une suite de mots arabes, le modèle p-contexte permet de calculer les probabilités suivantes :

Soit  $p$  ( $1 \leq p \leq n$ ), Pour  $j$  et pour tout  $j_1, \dots, j_p \in \{1, \dots, j-1, j+1, \dots, n\}$

$$\Pr(w_{i_j}, w_{i_{j_1}}, \dots, w_{i_{j_p}} \text{ se trouvent dans le même contexte})$$

On note par la suite cette probabilité :  $\Pr(w_{i_j}, w_{i_{j_1}}, \dots, w_{i_{j_p}} / \text{contexte})$

Ce modèle a été utilisé avec un autre modèle de syntaxe pour générer automatiquement des phrases arabes [Aouragh et al, 2006].

### 1.1 Utilisation du modèle p-contexte dans la génération des phrases arabes

La génération d'une phrase  $w_{i_1} w_{i_2} \dots w_{i_n}$  suppose les deux conditions suivantes :

- i) Chaque mot  $w_{i_j}$   $j \in \{1, \dots, n\}$  doit apparaître dans un contexte de taille  $p$  ( $1 \leq p \leq n$ ) avec tous les mots restants.
- ii) Il existe un chemin optimal  $s_{i_1}^*, s_{i_2}^*, \dots, s_{i_n}^*$  d'étiquettes de types syntaxiques des mots  $w_{i_1}, w_{i_2}, \dots, w_{i_n}$  tel que :

$$\Pr(w_{i_1}, \dots, w_{i_n}, s_{i_1}^*, \dots, s_{i_n}^*) \neq 0.$$

### 1.2 Conséquences

- 1) La condition (i) implique que pour tout  $j$  et pour tout  $j_1, \dots, j_p \in \{1, \dots, j-1, j+1, \dots, n\}$  on a :

$$\Pr(w_{i_j}, w_{i_{j_1}}, \dots, w_{i_{j_p}} \text{ se trouvent dans le même contexte}) \neq 0$$

On note par la suite cette probabilité :  $\Pr(w_{i_j}, w_{i_{j_1}}, \dots, w_{i_{j_p}} / \text{contexte}) \neq 0$

Ceci est équivalent à :

$$\prod_{j=1}^n \prod_{j_1=j+1}^n \prod_{j_2=j_1+1}^n \dots \prod_{j_p=j_{p-1}+1}^n \Pr(w_{i_j}, w_{i_{j_1}}, \dots, w_{i_{j_p}} / \text{contexte}) \neq 0 \quad (1)$$

- 2) La condition (ii) permet de vérifier si l'ordre des mots  $w_{i_1}, w_{i_2}, \dots, w_{i_n}$  est juste ou non.

Ceci est réalisé en se basant sur des connaissances grammaticales des mots  $w_{i_1}, w_{i_2}, \dots, w_{i_n}$ .

Nous avons modélisé cette condition par l'existence d'un chemin optimal  $s_{i_1}^*, s_{i_2}^*, \dots, s_{i_n}^*$  d'étiquettes de types syntaxiques des mots  $w_{i_1}, w_{i_2}, \dots, w_{i_n}$  tel que :

$$\Pr(w_{i_1}, \dots, w_{i_n}, s_{i_1}^*, \dots, s_{i_n}^*) \neq 0 \quad (2)$$

- 3) La synthèse des deux conditions donne la probabilité de générer la phrase  $w_{i_1} w_{i_2} \dots w_{i_n}$  comme étant le produit entre des probabilités (1) et (2) (les deux niveaux sémantique et syntaxique sont supposés indépendants) :

$$\Pr(w_{i_1}, \dots, w_{i_n}) = \beta_{d_n}^* \times \prod_{j=1}^n \prod_{j_1=j+1}^n \prod_{j_2=j_1+1}^n \dots \prod_{j_p=j_{p-1}+1}^n \Pr(w_{i_j}, w_{i_{j_1}}, \dots, w_{i_{j_p}} / \text{contexte}) \quad (3)$$

$$\times \Pr(w_{i_1}, \dots, w_{i_n}, s_{d_1}^*, \dots, s_{d_n}^*)$$

Où:  $\beta_{d_n}^* = \Pr(s_{d_n} \text{ soit état finale})$

## 2 Modèles à base de classes

L'utilisation du modèle p-gram et p-contexte est confrontée aux problèmes suivants :

- Le nombre de paramètres à estimer est élevé,
- L'estimation des paramètres nécessite un corpus d'apprentissage très large,
- Le problème des événements non représentés dans le corpus d'apprentissage (mots inconnus, etc).

Pour remédier à ces problèmes, on utilise la notion de classes regroupant des mots ayant un comportement similaire.



Quatre types de classes sont couramment utilisés :

- Classe syntaxique : elle regroupe les mots ayant la même catégorie grammaticale,
- Classe morphologique : elle regroupe les mots selon leur racine (lemme),
- Classe obtenue par la classification automatique,
- Classe sémantique : elle regroupe les mots ayant une relation sémantique.

L'utilisation des classes comporte au moins deux avantages :

- Le nombre d'événements à modéliser est moindre, ce qui nécessite moins de données d'apprentissage.
- L'utilisation des classes permet d'établir une généralisation de certains événements non vus dans le corpus d'apprentissage.

Dans le cas d'utilisation des classes syntaxiques, l'approximation de la probabilité  $P(w_t / w_{t-2} w_{t-1})$  de modèle trigram utilisée [Cerf-Danon et El-Bèze, 1991] est :

$$P(w_t / w_{t-2} w_{t-1}) \approx \sum_{c_t \in C} P(w_t / c_t) \times P(c_t / c_{t-2} c_{t-1}) \quad (4)$$

Avec :

$c(w_t) = c_t$  est la classe syntaxique du mot  $w_t$  et  $C$  est l'ensemble des classes syntaxiques.

### 3 Modèle p-contexte de classe

Dans ce travail nous avons adapté le modèle de langage n-classe à notre modèle p-contexte, la classification syntaxique est déjà réalisée dans le modèle syntaxique avec le modèle p-contexte pour la génération des phrases [Aouragh et al, 2006], c'est pourquoi nous nous sommes basés sur la classification sémantique des mots pour construire notre modèle p-contexte de classe.

Si on prend une phrase  $ph = w_{i_1}, \dots, w_{i_n}$ , la classe sémantique du mot  $w_{i_k}$ ,  $1 \leq k \leq n$  est définie par :

$$C(w_{i_k}) = C_{i_k} = \{w \in V / w_{i_1} w_{i_2} \dots w_{i_{k-1}} w w_{i_{k+1}} \dots w_{i_n} \text{ est une phrase juste}\}$$

Soit  $\zeta$  l'ensemble des classes sémantiques associées au vocabulaire du système, cette classification nous a permis d'adapter heuristiquement l'approximation de la probabilité  $P(w_i / h_i)$  du modèle n-gram [Bahl et al. 1978] (formule 4) à notre modèle p-contexte. Cette approximation est donnée par la formule suivante :

$$P(w_{i_1}, \dots, w_{i_n} / \text{contexte}) \approx \sum_{c_{i_1}, \dots, c_{i_n} \in C} P(w_{i_1} / c_{i_1}) \times \dots \times P(w_{i_n} / c_{i_n}) P(c_{i_1}, \dots, c_{i_n} / \text{contexte}) \quad (5)$$

### 4 Application

Comme application de ce modèle, nous avons pris le cas de  $p = 1$  (modèle bi-contexte). Dans ce cas la formule (5) devient :

$$P(w_{i_1}, w_{i_2} / \text{contexte}) \approx \sum_{c_{i_1}, c_{i_2} \in C} P(w_{i_1} / c_{i_1}) \times P(w_{i_2} / c_{i_2}) P(c_{i_1}, c_{i_2} / \text{contexte}) \quad (6)$$

La construction d'un modèle bi-contexte de classes consiste à estimer les paramètres suivants:

- $W = (P(w_t / c_i))_{1 \leq t \leq N, 1 \leq i \leq M}$  la matrice des probabilités d'apparition d'un mot  $w_t$  dans une classe sémantique  $c_i$ .
- $C = (P(c_i, c_j / \text{contexte}))_{1 \leq i \leq M, 1 \leq j \leq M}$  La matrice des probabilités que deux classes  $c_i$  et  $c_j$  se trouvent dans le même contexte.

#### 4.1 Estimation des paramètres

L'apprentissage d'un modèle p-contexte de classes consiste à estimer l'ensemble de ses probabilités à partir d'un corpus d'apprentissage. Si on prend un ensemble d'apprentissage  $R = \{ph_1, \dots, ph_K\}$  constitué par K phrases arabes,  $V = \{w_1, w_2, \dots, w_N\}$  le vocabulaire de notre système et  $\zeta = \{c_1, c_2, \dots, c_M\}$  l'ensemble de toutes les classes sémantiques associées au vocabulaire  $V$ .

L'utilisation de ce modèle nécessite seulement l'estimation de  $NM + \frac{1}{2}(M^2 + M)$  paramètres par rapport à l'estimation de  $\frac{1}{2}(N^2 + N)$  paramètres pour le modèle bi-contexte ( $M \ll N$ ).

#### 4.2 Expérimentation

Nous avons construit un corpus d'apprentissage contenant 213 phrases avec un vocabulaire de 590 mots regroupés dans 119 classes sémantiques, ces phrases sont étiquetées par 83 étiquettes de type syntaxique. L'évaluation de ce nouveau modèle est réalisée par un programme écrit en langage Perl, contenant deux modules :

- Module d'apprentissage : il permet d'estimer l'ensemble  $\{W, C\}$  des paramètres de notre modèle.
- Module de génération des phrases : il permet de générer des phrases à partir du vocabulaire.

#### 4.3 Résultats

Pour évaluer notre modèle, nous avons généré toutes les phrases possibles de quatre mots ayant la probabilité de génération non nulle (formule 3).

Le taux d'erreur utilisé dans notre travail est défini comme étant le pourcentage des phrases fausses générées par rapport à toutes les phrases générées par le système.

Le taux d'erreur exacte sur ces phrases est le suivant :

	bi-contexte	bi-contexte de classe
Nombre de phrases générées	255	436
Taux d'erreur	15.90%	22.80%

On remarque que le nombre de phrases générées par le modèle bi-contexte de classe est plus grand que celui de modèle bi-contexte, mais le taux d'erreur est élevé par rapport à ce dernier et ceci est logique car il y a des pertes d'information au niveau de classes. L'utilisation de ce modèle apporte une amélioration sur le temps d'exécution car il diminue le nombre de paramètres du système, de même il rend le modèle de génération un peu indépendant du vocabulaire.

Un apprentissage du modèle p-contexte de classe sur un corpus assez large, ainsi l'augmentation du nombre de classes sémantiques améliorera sans doute les performances de notre modèle de génération des phrases.

### 5 Conclusion

Les résultats de l'approximation de modèle p-contexte sont en général encourageants, cet approximation nous a permis de remédier aux problèmes cités avant. Le nombre de phrases générées par le modèle bi-contexte de classe est plus grand que celui de modèle bi-contexte, mais le taux d'erreur est élevé par rapport à ce dernier et ceci est logique car il y a des pertes d'information au niveau de classification.

Un apprentissage du modèle p-contexte de classe sur un corpus assez large et l'augmentation de l'ordre du modèle p-contexte éliminera sans doute plusieurs erreurs lors de la génération des phrases.

## **References**

1. L. AOURAGH, J. ALLAL et A. YOUSFI, 2006, A Stochastic Language Model for Automatic Generation of Arabic Sentences, *Georgian Electronic Scientific Journal : Computer Science and Telecommunications*.
2. L. AOURAGH, J. ALLAL et A. YOUSFI, JIMIO'1 2005, Modèle stochastique de langage pour la génération automatique des phrases en arabe, *Les premières Journées Internationales de Mathématiques & Informatique*, université Mohammed 1er, Faculté des Sciences, Oujda, Maroc.
3. Aouragh E., Yousfi A., Jelloul A., INFORSID'2006, Atelier SIA (Systèmes d'Information Arabisés). « La modélisation des niveaux syntaxique et sémantique pour la génération automatique des phrases arabes ». Hammamet, Tunisie, 31 mai 2006.
4. H. CERF-DANON et M. EL-BÈZE. 1991. Three different probabilistic language models : comparison and combination. Pages 297–300 of : *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. Toronto, Canada.
5. M. EL-BÈZE et A.-M. DEROUAULT. 1990. A morphological model for large vocabulary speech recognition. Pages 577–580 of : *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. Albuquerque, New Mexico, USA
6. BAHL L.R., BAKER J.K., COHEN P.S., JELINEK F., LEWIS B.L., et MERCER R.L., 1978, Recognition of a continuously read natural corpus. In : *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Tulsa.

## English-French Cross Language Text Categorization

Bentaallah Mohamed Amine<sup>1</sup> and Malki Mimoun<sup>1</sup>

<sup>1</sup> EEDIS Laboratory, Departement of computer science  
Sidi bel abbes, 22000. ALGERIA  
{mabentaallah, malki}@univ-sba.dz

**Abstract.** This article is essentially dedicated to the problem of Cross-Language Text Categorization that consists of using a labelled dataset in one language (L1) to classify unlabelled data in other language (L2). This problem can be formulated as: "how can we train a text classifier for one language using labelled examples for another language?" The proposed approach is based on the idea to spread the utilization of WordNet in Text Categorization towards Cross-Language Text Categorization. Experimental results of the English-French Cross-Language Text Categorization of the Multilingual Reuters corpus show that the idea we describe are promising and deserve further investigation.

**Keywords:** Multilingual, Text Categorization, WordNet, Multilingual Reuters Corpus

### 1 Introduction

Text Categorization refers to the task of automatically assigning one or multiple predefined category labels to free text documents. It consists of assigning a Boolean value to each pair  $\langle d_j, c_i \rangle$ , where  $d_j$  is a document and  $c_i$  is a category. A value of T(True) assigned indicates a decision to file  $d_j$  under  $c_i$ , while a value of F(False) indicates a decision not to file  $d_j$  under  $c_i$  [1].

With the rapid emergence and proliferation of Internet and the trend of globalization, a tremendous number of textual documents written in different languages are electronically accessible online. Efficiently and effectively managing these textual documents written in different languages is essential to organizations and individuals. In this article, we propose a new approach for Cross-Language Text Categorization that consists in spreading the use of WordNet in text categorization to categorize documents coming from different languages. The proposed approach is based on the translation of the unlabelled documents towards the language of Shakespeare in order to be able to use the WordNet ontology thereafter.

The paper is organized as follows. In section 2, a definition of Multilingual Text Categorization is presented. In Section 3, we briefly review some related works for Multilingual Text Categorization. We describe the two proposed approaches with all stages in section 4. Experimental results are reported in section 5. Finally, conclusion and future works are reported in section 6.

### 2 Cross-Language Text Categorization

The Cross-Language Text Categorization is a new area in text categorization in which we have to cope with two or more languages (e.g English, Spanish and Italian). In Cross-Language Text Categorization, three scenarios can be distinguished:

**-Poly-lingual training:** In this scenario, the system is trained using training examples from all the different languages. A single classifier is build using a set of labelled train documents in all languages, which will classify documents from different languages. This scenario excluded the use of translation strategies; therefore, no distortion of information or loss is committed.

**-Cross-lingual training:** The system uses labelled training for only one language to classify documents in other languages. This approach is what we are interested in this paper. To solve this problem, we can use the translation in different ways:

- **Training-Set translation:** In this approach, the labelled set is translated into the target language which then is used to train a classifier for this language. So, the Cross-lingual training became a Poly-lingual training.
- **Test-Set translation:** This approach consists on translating the unlabelled documents into one language ( $L1$ ). To classify the unlabelled translated documents, the system is trained using the labelled training set for language ( $L1$ ). So, the Cross-Language Text Categorization became Monolingual.

**-Esperanto language:** This approach uses a universal reference language which all documents are translated to. This universal language should contain all properties of the languages of interest and be organized in a semantic way.

### 3 Related Work

When we embarked on this line of research, we have note a lack on works addressing directly the area of Cross-Language Text Categorization. The majority of research works essentially comes of the Cross-Language Text Retrieval. Indeed, the two areas are based on the same aspects (similarity between texts, comparison of documents with queries or class profiles).

R.Jalam et al. presented in [2] three approaches for Multilingual Text Categorization that are based on the translation of documents toward a language of reference. The authors claimed having got good enough results.

A.Gliozzo and C.Strapparava propose in [3] a new approach to solve the Multilingual Text Categorization problem based on acquiring Multilingual Domain Models from comparable corpora to define a generalized similarity function (i.e. a kernel function) among documents in different languages, which is used inside a Support Vector Machines classification framework. The results show that the approach largely outperforms a baseline.

These last years, researches showed that using ontologies in monolingual text categorization is a promising track. J.Guyot proposed in [4] a new approach that consists in using a multilingual ontology for Information Retrieval, without using any translation. He tried only to prove the feasibility of the approach. Nevertheless, it still has some limits because the used ontology is incomplete and dirty.

Intelligent methods for enabling concept-based hierarchical Multilingual Text Categorization using neural networks are proposed in [5]. These methods are based on encapsulating the semantic knowledge of the relationship between all multilingual terms and concepts in a universal concept space and on using a hierarchical clustering algorithm to generate a set of concept-based multilingual document categories, which acts as the hierarchical backbone of a browseable multilingual document directory. The concept-based multilingual text classifier is developed using a three-layer feed-forward neural network to facilitate the concept-based Multilingual Text Categorization.

### 4 Our Approach

The suggested approach is composed of two phases. The first relates on the training phase that consists on creating conceptual categories profiles. These conceptual profiles will contain concepts that characterize best a category with regard to the other categories. The second phase is the classification, it consists on using machine translation techniques to translate documents into English language in order to generate its conceptual vector using WordNet. After this, we have to weight document vector and categories profiles, then to calculate distances between them in order to be able to decide the adherence of the document to a category or an another one.

#### 4.1 Training phase

The first issue that needs to be addressed in this phase is: how to represent texts so as to facilitate machine manipulation but also to retain as much information as needed?. The commonly used text representation is the Bag-Of-Words, which simply uses a set of words with their number of occurrences to represent documents and categories [6]. This representation was disadvantaged by the ignorance of any relation between words, thus learning algorithms are restricted to detect patterns in the used terminology only, while conceptual patterns remain ignored. In our approach, the training phase consists on using WordNet to create profiles categories which will contain concepts (synsets in WordNet) that characterize best one category with regard to the other categories. For this purpose, three steps are required:

**Mapping terms into synsets** The most straightforward representation of documents relies on term vectors. But, in some cases, terms with different concepts can even be replaced with only one higher level concept without negative effect on performance of the classifier. Obviously, mapping terms to synsets is an effective and reasonable method to reduce the dimensionality of the vector space.

In the most case, one word may have several meanings and thus one word may be mapped into several synsets which may add noise to the representation and may induce a loss of information. In this case, we need to determine which meaning is being used, which is the problem of sense disambiguation [7]. In our approach, we used a simple disambiguation strategy that consists on considering only the most common meaning of the term (first ranked element) as the most appropriate. So our mapping process consists on replacing each term by its most common meaning.

**Capturing Relationships between Synsets** After mapping terms into synsets, this step consists on using the WordNet hierarchies to capture some useful relationships between synsets (hypernymy in our case).

**Creating conceptual categories profiles** Selection methods for dimensionality reduction take as input a set of features and output a subset of these features, which are relevant for discriminating among categories [8]. For our approach, a feature selection technique is necessary in order to reduce the big dimensionality by creating the conceptual categories profiles. For this purpose we have used the  $\chi^2$  multivariate statistic for feature selection that consists on extracting, for each category, the  $K$  better synsets (our features) characterizing best the category compared to the others. With this intention, the matrix (synsets-categories) representing the total number of occurrences of the  $p$  synsets in the  $m$  categories is calculated. The total sum of the occurrences is noted  $N$ . The values  $N_{jk}$  represent the frequency of the synset  $s_j$  in the category  $c_k$ . Then, contributions of these synsets in discriminating categories are calculated as indicated in the following equation, and then sorted by descending order for each category.

$$C_{jk}^{\chi^2} = N \frac{(f_{jk} - f_{j.}f_{.k})^2}{f_{j.}f_{.k}} \times \text{sign}(f_{jk} - f_{j.}f_{.k}) \quad \dots\dots (1)$$

#### 4.2 Classification phase

The classification phase consists on using the conceptual categories profiles in classifying unlabelled documents in different languages. Our classification phase consists on:

**Translation and generation of the conceptual vector** The translation of the text to be classified in the language of training corpus is also a paramount stage. The objective here is not to produce a translated text accurately recalling the semantic properties of the original text, but to provide a text ensuring a sufficient quality of classification. It is obvious that the result obtained will depend on the translator used. For that, we used JWT<sup>1</sup>(Java Web Translator) library which provides automatic language translation for 14 languages including English, Spanish, French, Italian, Deutsch, Greek, Chinese, Japanese, Russian, ect. After translating document, we have to use WordNet in order to generate a conceptual vector for the document (Mapping terms into synsets and Capturing relationships between synsets).

<sup>1</sup> This package is available on: <http://sourceforge.net/projects/webtranslator>

**Weighting** This stage consists on weighting conceptual categories profiles and conceptual vector of the unlabelled document. Each weight  $w(c; s)$  expresses the importance of synset  $s$  in vector of  $c$  with respect to its frequency in all training documents. The objective of using a feature weight rather than plain frequencies is to enhance classification effectiveness. In our experiments, we used the standard *tfidf* function, defined as:

$$tfidf(s_k, c_i) = tf(s_k, c_i) \times \log\left(\frac{|C|}{df(s_k)}\right) \quad \dots\dots(2)$$

**Distance computation** The similarity measure is used to determine the degree of resemblance between two vectors. To achieve reasonable classification results, a similarity measure should generally respond with larger values to documents that belong to the same class and with smaller values otherwise. The dominant similarity measure in information retrieval and text classification is the cosine similarity between two vectors. It can be calculated as the normalized dot product:

$$S_{i,j} = \frac{\sum_{s \in i-j} tfidf(s,i) \times tfidf(s,j)}{\sqrt{\sum_{s \in i} tfidf^2(s,i) \times \sum_{s \in j} tfidf^2(s,j)}} \quad \dots\dots(3)$$

## 5 Experimental Results

### 5.1 Dataset for evaluation

For our experimentations, we extracted a bilingual dataset from Reuters Corpus Vol. 1 and 2(RCV1, RCV2) using English training documents (RCV1) and French test documents (RCV2). Our dataset is based on topic (category) codes with a rather varying number of documents per category as shown in Table1.

**Table 2.** The 8 used categories of the Multilingual Reuters Corpus..

Code Category	Description Category	#English	#French
GREL	Religion	401	121
C183	Privatisations	401	401
GSPO	Sport	401	401
GFAS	Fashion	124	39
GSCI	Science and technology	401	401
GJOB	Labour Issues	401	401
GTOUT	Travel and Tourism	394	65
GVIO	War, Civil war	401	401

### 5.2 Evaluation method

Experimental results reported in this section are based on the so-called “*F1* measure”, which is the harmonic mean of precision and recall.

$$F_1(i) = \frac{2 \times precision \times recall}{precision + recall} \quad \dots\dots(4)$$

In the above formula, precision and recall are two standard measures widely used in text categorization literature to evaluate the algorithms effectiveness on a given category. We also use the macroaveraged *F1* to evaluate the overall performance of our approach on given datasets. The macroaveraged *F1* computes the *F1* values for each category and then takes the average over the per-category *F1* scores. Given a training dataset with  $m$  categories, assuming the *F1* value for the  $i$ -th category is  $F_1(i)$ , the macroaveraged *F1* is defined as :

$$MacroAveragedF_1 = \frac{\sum_{i=1}^m F_1(i)}{m} \quad \dots\dots(5)$$

### 5.3 Results

In order to be able to show the utility of the use of WordNet in Cross-Language Text Categorization, we tested the suggested approach on our multilingual dataset. The results of the experimentations are presented in Table 2.

**Table 2.** MacroAveraged *F1* results on the used dataset.

Profiles size	MacroAveraged
K=100	0.703
K=200	0.705
K=300	0.721
K=400	0.722
K=500	0.722
K=600	0.729
K=700	0.738
K=800	0.746
K=900	0.749

Concerning the profiles size, it is noted that for the two datasets, the best performances are obtained with size profile  $k = 900$ . Indeed, the performances improve more and more by increasing the size of profiles. A MacroAveraged *F1* value of 0.749 proves the utility of the proposed approach.

## 6 Conclusion

In this paper, we proposed a new approach for Cross-Language Text Categorization which is based on the one hand on the use of WordNet and other hand on the use of machine translation techniques. The obtained results are encouraging and carry out us to confirm that the use of WordNet in Cross-Language Text Categorization is a promising track.

Our future works will concern the use of WordNet in distance computation in order to be able to test the use of the semantic distances in stead of the statistics distances. Another track consists on using other disambiguation, selection and Weighting techniques.

## References

1. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys*. (2002) 1–47.
2. Jalam, R., Clesh, J., Rakotomalala, R.: Cadre pour la catégorisation de textes multilingues. 7èmes Journées internationales d'Analyse statistique des Données Textuelles. Louvain-la-Neuve, Belgique (2004) 650–660.
3. Gliozzo, A.M., Strapparava, C.: Cross Language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Ann Arbor, Michigan, USA (2005) 9–16
4. Guyot, J., Radhouani, S., Falquet, G.: Ontology-based multilingual information retrieval. In *CLEF Workhop, Working Notes Multilingual Track*, Vienna, Austria (2005) 21–23
5. Chau, R., Yeh, C.H., Smith, K.: A Neural Network Model for Hierarchical Multilingual Text Categorization. In *proceeding of ISSN-05 Second International Symposium on Neural Networks*, Chongqing, China (2005) 238–245
6. Peng, X., Choi, B.: Document Classifications Based On Word Semantic Hierarchies. In *proceedings of the IASTED International Conference on Artificial Intelligence and Applications*. (2005) 362–367.
7. Ide, N., Vieu, J.: Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*. (1998).
8. Dash, M., Liu, H.: Feature selection for classification. *Intelligent Data Analysis*. (1997).



## Lexicon KalimNet : Calcul Automatique De La Synonymie A Partir D'Un Dictionnaire Monolingue Arabe

Zouaoui Amine AEK<sup>1</sup>, Lehireche Ahmed<sup>1</sup>, Malki Mimoun<sup>1</sup>

<sup>1</sup>Université Djillali Liabes de Sidi Bel abbes,  
EEDSI, département d'informatique, Algérie.  
{[z.a.amine@gmail.com](mailto:z.a.amine@gmail.com), [elhir@yahoo.com](mailto:elhir@yahoo.com), [malki-m@yahoo.fr](mailto:malki-m@yahoo.fr)}.

**Résumé.** Avec le développement qu'a connu le traitement automatique du langage naturel ces dernières décennies ainsi que ses applications correspondantes comme la recherche d'informations, l'extraction d'informations ou la traduction automatique, la disponibilité des lexicons informatiques pour les différentes langues est devenue une nécessité primordiale. Dans ce contexte, le laboratoire EEDISL – département d'informatique, UDL de Sidi bel abbés - a récemment proposé et a initié le projet KalimNet, qui a pour but la création d'un lexicon informatique pour la description de la sémantique lexicale arabe. Notre intervention dans ce projet s'est focalisée sur une seule catégorie syntaxique qui est les verbes arabes et l'on s'est fixé comme objectif la création de l'ensemble des concepts en utilisant les synsets comme modèle de représentation du sens. Notre approche utilise un graphe construit à partir des définitions d'un dictionnaire monolingue de la langue arabe contemporaine. L'application d'une mesure pour le calcul de similarité entre les sommets (verbes) du graphe construit, nous permettra de détecter les verbes similaires et d'identifier ainsi pour un verbe donné l'ensemble de ces synonymes.

**Mots clés :** Traitement automatique du langage, Lexicon informatique, machine readable dictionary, sémantique lexicale.

### 1 Introduction

Tout système de traitement automatique du langage naturel nécessite plusieurs ressources dans son processus d'analyse de texte écrit ou parlé dans un langage humain particulier, ainsi au niveau lexical du système [1], le lexicon peut constituer une ressource indispensable, ce dernier fournira des informations sur le lexique du langage et sa nature dépend de l'ensemble d'informations présentées, ainsi un lexicon peut être assez simple en décrivant seulement la catégorie syntaxique de chaque mot ou plus complexe et pourra donc décrire pour chaque mot contenu, la classe sémantique à laquelle il appartient, l'ensemble d'arguments possibles, le sens de ses définitions selon le modèle adopté par le système pour la représentation du sens ainsi que d'autres informations.

L'algorithme choisi est celui proposé par V.blondel et P.senellart [4], pour le calcul de similarité entre les sommets de deux graphes orientés [5], qui a prouvé que son application est fortement intéressante dans le cas de recherche de mots similaires au sens de la synonymie en utilisation un dictionnaire monolingue comme source lexicale.

L'algorithme proposé exige, un graphe devant être construit à partir du dictionnaire choisi, la construction du graphe doit se faire de la manière suivante : Chaque entrée (mot) du dictionnaire est un sommet du graphe et un arc relie un sommet  $i$  à  $j$  si le mot  $j$  apparaît dans la définition de  $i$ .

Cet algorithme n'a besoin pour calculer les synonymes pour un mot  $w$  que d'une partie du graphe. Ce sous-graphe est appelé graphe de voisinage et il est construit à partir des sommets qui sont pointés par  $w$  et de ceux qui pointent vers  $w$ . Les détails de l'algorithme et du calcul de la similarité seront présentés dans la section 3.

Suite au choix de l'algorithme reste à définir le type de dictionnaire qu'on devra utiliser, ce choix est argumenté dans la section 4 et la méthodologie de construction du graphe y est présentée.

## 2 Travaux similaires

On présente dans cette section quelques travaux concernant le calcul automatique de mots similaires ou synonymes à partir de dictionnaire ou de large corpus textuel.

### 2.1 La méthode des Distances

Cette méthode est basée sur l'hypothèse que deux mots sont synonymes que s'ils paraissent dans la définition des mêmes mots et ont les mêmes mots dans leurs définitions. L'une des possibilités de formaliser cette condition est de définir une mesure de distance entre deux mots  $i$  et  $j$  : on compte le nombre de mots qui apparaissent seulement dans l'une des définitions mais pas dans les deux et on ajoute à cela le nombre de mots qui utilisent soit  $i$  ou  $j$  dans leurs définitions, mais pas les deux à la fois. En considérant le graphe complet extrait d'un dictionnaire et sa matrice d'adjacence, en prenant  $i$  et  $j$  deux sommets du graphe, l'expression de la distance entre les deux sommets  $i$  et  $j$  est ainsi formulé :

$$d(i,j) = \|(A_{i,} - A_{j,})\| + \|(A_{,i} - A_{,j})^T\|.$$

$\| \cdot \|$  est la norme L1 du vecteur (la somme des magnitudes de toutes les entrées). Donc pour un mot  $i$  donné, on peut calculer toutes les distances  $d(i,j)$  puis trier les mots selon leurs distances par ordre croissant. Cette méthode présente la particularité d'être applicable sur la totalité d'un graphe d'un dictionnaire mais sa sera au détriment de la précision des résultats dans l'éventualité où un graphe de voisinage donnerait de meilleurs résultats.

### 2.2 Analyse syntaxique légère

C'est une méthode basée sur le principe de *contexte syntaxique*, elles s'appuie sur le contexte grammatical qui peut être représenté par le tuple  $(w,r,w')$  où  $w$  et  $w'$  sont deux mots et  $r$  caractérise la relation entre  $w$  et  $w'$  : par exemple, deux noms sont similaires s'ils paraissent comme sujets ou objet direct d'un même verbe.

Parmi les travaux illustrant cette technique, on distingue celui de Grefenstette [2] qui a proposé l'approche SEXTANT opérant sur de large corpus de texte afin de n'extraire que les noms similaires. Le système SEXTANT utilise un analyseur lexical pour l'extraction des unités lexicales ainsi qu'un détecteur de phrases nominales et verbales en définissant des règles de début, de fin et de continuation.

Six contextes syntaxiques ont été défini par Grefenstette à travers des relations, le système analysera les phrase délimitées auparavant pour extraire l'une des relations syntaxiques suivantes :

<b>ADJ</b>	Un adjectif modifiant un nom
<b>NN</b>	Un nom modifiant un nom
<b>NNPREP</b>	Un nom qui est l'objet d'une proposition modifiant un mot
<b>SUBJ</b>	Un nom qui est sujet d'un verbe
<b>DOBJ</b>	Un nom qui est l'objet direct d'un verbe
<b>IOBJ</b>	Un Nom dans une phrase prépositionnelle modifiant un verbe

Après la phase d'analyse, chaque mot lui sera associé un certain nombre d'attributs : Représentant les mots qui le modifient tout en précisant le type de la relation syntaxique. Le calcul de similarité entre mots est défini ainsi :

Chaque attribut lui est assigné un poids selon la formule :

$$poid(att) = 1 + \sum_{nom \in} \frac{P_{att,nom} \log(P_{att,nom})}{\log(Nombre\ total\ des\ relation)}$$

Où :



$$P_{att_i} = \frac{\text{Nombre de fois att apparait au}}{\text{Nombre total des attribut de}}$$

La similarité de mesure utilisée par Grefenstette est la mesure de similarité pondéré de Jaccard qui est défini par :

$$jac(i, j) = \frac{\sum_{att \text{ attribut de } i \text{ et de } j} poids(att)}{\sum_{att \text{ attribut de soit } i \text{ ou soit } j} poids(att)}$$

### 3 Algorithme de calcul des synonymes

Cette section est consacrée à la présentation de l'algorithme de calcul des synonymes en vue de l'appliquer à un dictionnaire de la langue arabe. L'algorithme choisi est conçu sur la base de l'hypothèse suivante : *les mots synonymes ont plusieurs mots en communs dans leur définition et sont utilisés dans les définitions de plusieurs mots en communs.*

Pour mettre en œuvre une méthode permettant de vérifier la condition avancée par l'hypothèse ci-dessus, on aurait besoin du graphe de voisinage  associé à un mot w, assorti d'un algorithme de calcul de similarité capable d'identifier les mots du graphe  qui sont les plus similaires à w, c'est-à-dire ceux qui ont le plus de mots en commun et qui sont pointus par ou pointent vers w.


Considérons le graphe formé des trois sommets 1, 2 et 3, qu'on notera  ayant la structure suivante :



Fig. 1. Structure de graphe à trois sommets

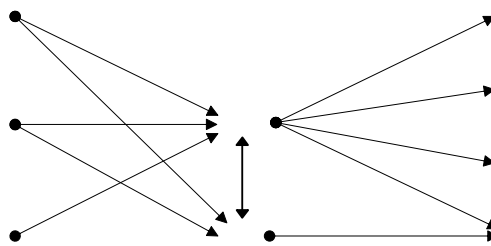




Fig. 2.: sous graphe du graphe de voisinage du mot **بحث** .

Le graphe de la figure 2 présente les similarités suivantes avec le graphe  : { **ياث**, **نيث**, **جس** } sont similaires au sommet 1, { **اجتهد**, **سال**, **استقصى**, **فتش** } sont similaires au sommet 3 et { **بحث**, **حفر** } sont similaires au sommet 2. On dira donc qu'un sommet est similaire au sommet 2 du graphe  s'il pointe vers des sommets similaires au sommet 3 et s'il est pointu par des sommets similaires au sommet 1.

En se référant à l'hypothèse avancée pour la définition des synonymes, la recherche des mots synonymes du mot  $w$  est traduite par la recherche dans le graphe de voisinage  $G(w)$ , des sommets qui sont les *plus similaire* au sommet 2 du graphe  $G(w)$ . Ainsi la matrice de similarité associée à  $G(w)$  et  $G(v)$  sera utilisée.

Finalement l'algorithme de calcul de similarité à partir de  $G(w)$  et  $G(v)$  est défini ainsi :

1. Associer trois scores à chaque sommet de  $G(w)$  :  $s_1$ ,  $s_2$  et  $s_3$  et les initialiser à la valeur 1.
2. En considérant  $i$  un sommet du graphe  $G(w)$  et  $v$  le vecteur formé à partir des scores  $s_1, s_2, s_3$ , Itérer les calculs suivants pour chaque sommet  $i$  de  $G(w)$  :

$$a. \begin{cases} x_1^i = \sum_{j \in N(i)} x_1^j \\ x_2^i = \sum_{j \in N(i)} x_2^j + \sum_{j \in N(i)} x_3^j \\ x_3^i = \sum_{j \in N(i)} x_3^j \end{cases}$$

$$b. x_k = x_k / \|x_k\|, k = 1, 2, 3 \quad (\| \cdot \| \text{ est la norme euclidienne}).$$

3. S'arrêter une fois la convergence atteinte.

Après le calcul, la matrice de similarité sera constitué des trois vecteurs normalisés  $s_1, s_2$  et  $s_3$ . Le vecteur normalisé  $s_1$  qui contient donc les scores centraux et après avoir été réordonné en ordre décroissant, nous fournira une liste de synonymes classés du plus pertinent au moins pertinent.

#### 4 Le Dictionnaire source et construction du graphe

Un bon nombre de dictionnaires monolingues arabes sont devenu par le temps des références dans la linguistique arabe. Hormis le critère de la richesse et le type du vocabulaire défini ainsi que la rigueur et la richesse des informations lexicales contenues dans le dictionnaire, un autre critère important déterminera le choix de notre dictionnaire source.

Les anciens dictionnaires arabes tels que Lissan Al-'arab ou El-Kamous El-Mouhit présentent la particularité suivante ; informations morphologiques (formes dérivées, pluriel,...etc.), définition et exemple(s) d'utilisation(s) se trouvent souvent contiguës dans un seul paragraphe et sont donc reliées par l'utilisation des particules de la langue arabe (adverbes, conjonctions et prépositions). Il est donc difficile de localiser dans ce texte la définition propre (nettoyée des exemples et des informations morphologiques) d'un sens sans passer par une analyse syntaxique lourde voir même sémantique, ceci a pour effet de rendre la tâche de la construction du graphe complexe. Le choix s'est donc orienté vers l'utilisation de l'un des dictionnaires monolingues contemporains de la langue arabe qui ont la particularité d'utiliser un texte structuré pour la définition d'une entrée. *El-raïd*, *EL-mounjid* et *EL-Moujiz* sont des exemples devenus des références de tels dictionnaires.

Vu que les dictionnaires contemporains monolingues arabes cités ci-dessus ne sont pas disponibles sous format exploitable par la machine, on a préconisé l'utilisation de l'un des lexicons arabes mis en disposition en ligne par la société Sarkh [7], on y trouve donc deux dictionnaires contemporains répondants à nos besoins : "Al Rannye (الغني)", "Al Wassit (الوسيط)", chaque entrée du dictionnaire est décrite dans un fichier Html propre. Notre choix s'est porté sur le dictionnaire Al Rannye où l'ensemble des entrées de ce dictionnaire est décrit par près de 30.000 fichiers Html.

Le texte structuré utilisé dans la définition d'une entrée dans le dictionnaire choisi, nous permettra de segmenter le texte de la définition de chacun des sens et ainsi extraire la définition proprement dite (sans exemple d'utilisation et sans contexte).

Dans notre cas, seules les entrées verbales du dictionnaire nous intéressent. Le texte de la définition du sens d'une entrée verbale est structuré de la manière suivante :

- **Entrée\_verbale** :[Informations grammaticales.]
- Sens 1- [« Contexte » :] PV [« exemple d'utilisation »]\*
- .
- Sens n- [« Contexte » :] PV [« exemple d'utilisation »]\*

« Contexte » est une phrase qui contient le verbe défini, son rôle est de faire apparaître pour un sens donné, des contraintes sémantiques sur les éléments de la phrase qui accompagnent le verbe et peut concerner donc le sujet, les prépositions ou le complément d'objet de la phrase. Ainsi, la phrase du contexte peut préciser par exemple le type du sujet ou du complément: être vivant, animé/non-animé, machine, outil, personnes ou objets particuliers...etc.

« Exemple d'utilisation » est une phrase qui illustre une utilisation du verbe défini.

PV représente une phrase verbale (peut inclure des lettres de conjonctions) ou seulement un verbe.

Ainsi on préconise l'utilisation d'un analyseur morphologique arabe supportant les voyelles arabes, en occurrence celui de Buckwalter [8], pour détecter la catégorie lexicale de chaque mot analysé. L'utilisation de la version java de l'analyseur morphologique de Buckwalter intitulé Aramorph[10], a été prévu.

L'utilisation du texte arabe voyellé par le dictionnaire source et le support des voyelles arabes par l'analyseur morphologique sont deux conditions nécessaires pour la détection de la catégorie syntaxique d'un mot arabe de la façon la moins ambigu possible. Le tableau suivant montre les différentes possibilités pour voyeller un mot et les différents sens résultants avec la catégorie syntaxique correspondante :

Puisque seules les entrées verbales du dictionnaire seront traitées, le fait que le dictionnaire arabe "Al Rannye" indique précisément les entrées verbale comme telles, peut être considéré comme un réel avantage. Cette indication ainsi que d'autres informations supplémentaires sur le verbe, se trouve dans l'information grammaticale associé à l'entrée verbale : (استعدّ - [ع د د]. (ف: سدا. لازم، م. بحرف).

Ainsi les entrées verbales sont déduites directement sans passer par une analyse morphologique de toutes les entrées du dictionnaire.

La détection des verbes apparents dans la définition d'une entrée verbale se fera grâce à l'analyse des solutions générées par l'analyseur morphologique Aramorph.

Vu qu'il faut adapter le format du dictionnaire en entrée à un autre, qui sera exploitable par notre algorithme de calcul, on a préconisé la création d'une base donnée constituée de trois tables qui sont définies par les schémas relationnels suivant :

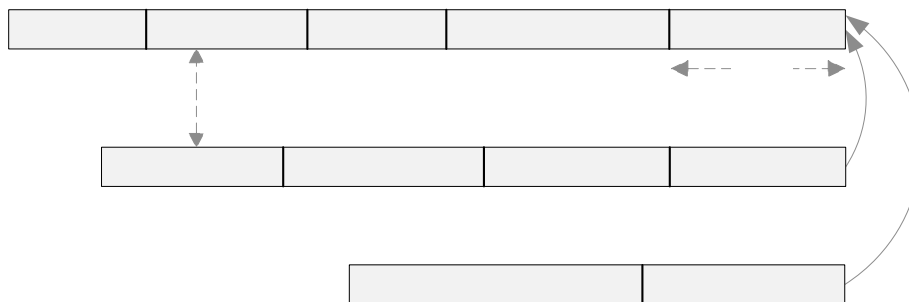


Fig. 3. schémas des tables de la BDD.

On décrira quelques attributs issus des différentes tables de la base de données :

- Racine, contient la racine trilatérale ou quadrilatérale de chaque entrée du dictionnaire.

- POS, contient l'information de la catégorie syntaxique (verbe, nom...etc.) pour chaque entrée du dictionnaire.
- N\_Mot de la table Définition est une clé étrangère référençant la clé N\_Mot de la table Entrée.
- Exemple, contiendra l'exemple d'utilisation pour chaque entrée du dictionnaire.
- Verbes\_Definissants contient tous les verbes (lemmatisés) apparents dans la définition pour chaque entrée du dictionnaire.

Les deux tables Entrée et Définition représenteront le dictionnaire source avec toutes les entrées qui y figure. Ainsi le dictionnaire sera retranscrit intégralement dans notre base de données.

La relation (Entrée-Verbe, Verbes-définissants) décrite dans la table Définissant, nous permettra de construire le graphe de voisinage associé à une entrée verbale *v* du dictionnaire.

Finalement un Parse approprié des fichiers Html du dictionnaire choisi, nous permettra d'extraire toutes les informations nécessaires à mettre dans les deux tables Entrée et Définition de notre Base de données.

## 5 Conclusion

On a présenté par le présent article, une approche permettant d'extraire les mots synonymes en utilisant un dictionnaire monolingue arabe comme seule source lexicale, elle s'avère donc être d'une grande utilité pour la reformulation des requêtes en langage naturel ou la compilation des dictionnaires de synonymes comme étant un support (très rare pour la langue arabe). Cette approche pourrait être très utile pour le projet KalimNet, dans la mesure où le modèle de représentation du sens serait les ensembles de synonymes (synset). Dans cette optique on s'est fixé comme objectif d'une part d'implémenter l'approche proposée au profit du projet KalimNet et d'autre part de réaliser une étude comparative entre les synsets générés et les synsets du lexicon Arabic WordNet.

## References

1. Liddy, E.D. 2001. *Natural Language Processing*. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
2. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, MA, 1994.
3. P. Senellart et V. D. Blondel, « *Automatic discovery of similar words* ». M. W. Berry et M. Castellanos, éditeurs, Survey of Text Mining II: Clustering, Classification and Retrieval, p. 25-44. Springer-Verlag, janvier 2008.
4. Vincent D. Blondel and Pierre P. Senellart *Automatic extraction of synonyms in a dictionary*,
5. Vincent, D. Blondel, Anah Gajardo, Maureen Heymans, Pierre Senellart, Paul Van Dooren . *A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching*, Society for Industrial and Applied Mathematics Vol. 46, No. 4, pp. 647-666, 2004.
6. Buckwalter, T (2002), *Arabic Morphological Analysis*, <http://www.quamus.org/morphology.htm>.
7. Sarkh Lexicons, <http://lexicons.sakhr.com/>.
8. Aramorph, <http://www.nongnu.org/aramorph/french/index.html>.
9. Doumi N. et A. Lehireche, *Une ontologie pour le lexique arabe*, in proceeding du 2ème congrès international de "l'ingénierie de la langue arabe et de l'ingénierie de la langue", CRSTDLA, UA, 2005.

## Utilisation des N-Grammes dans la Classification Automatique Non supervisée de Textes

Abdelmalek Amine<sup>1,2</sup>, Zakaria Elberrichi<sup>1</sup>, Michel Simonet<sup>3</sup>, Mimoun Malki<sup>1</sup>

<sup>1</sup> Laboratoire EEDIS, Département d'informatique  
Faculté de l'ingénieur. Université Djillali Liabes  
Sidi bel Abbes – Algérie

<sup>2</sup> Département d'informatique, Centre Universitaire Taher Moulay de Saida  
amine\_abdl@yahoo.fr, elberrichi@gmail.com, malki\_m@yahoo.com

<sup>3</sup> Laboratoire TIMC-IMAG  
INstitut de l'Ingénierie et de l'Information de Santé (IN3S)  
Faculté de médecine. Université Joseph Fourier  
Grenoble - France  
michel.simonet@imag.fr

**Résumé.** Dans cet article, nous proposons la méthode des cartes auto-organisatrices de Kohonen (SOM) pour la classification non supervisée de documents textuels basés sur des vecteurs centrés non pas sur les mots mais sur les n-grammes. Cette combinaison sera examinée dans plusieurs expériences utilisant la distance euclidienne et la distance du cosinus comme mesures de similarité et plusieurs valeurs de n. Les données utilisées dans nos expérimentations sont issues des textes du corpus de référence Reuters-21578. Pour évaluer les qualités des classifications obtenues nous utilisons la F-mesure et l'entropie.

**Mots clés:** Classification non supervisée, cartes auto-organisatrices de Kohonen, n-grammes, similarité, Reuters-21578.

### 1 Introduction

Avec l'abondance et la croissance rapide du nombre de documents présents sous forme numérique (Internet, bibliothèque numérique, CD-ROM,...), la catégorisation ou classification automatique de textes est devenue un domaine de recherche important.

La catégorisation ou classification automatique de textes est l'action de distribuer par catégories ou classes un ensemble de documents qui ont des caractéristiques communes.

Cependant le terme « classification » est ambiguë, il ne sépare pas entre l'action de création de classes (classer) et l'action d'affectation à une classe (classifier) qui sont deux approches différentes.

En terminologie anglaise le terme est non ambigu ; on parle de :

- « categorization » ou « classification » lorsqu'il s'agit d'affectation à une classe (les classes étant connues d'avance), dans ce cas on se situe dans le cadre de l'apprentissage supervisé.
- « clustering » (classification non supervisée) quand il s'agit de création de classes ou regroupement d'un certain nombre d'objets similaires sans a priori, on se situe alors dans le cadre de l'apprentissage non supervisé.

La classification non supervisée ou « clustering » est automatique ; elle fait émerger des classes latentes (cachées), non étiquetées. Les classes sont à découvrir automatiquement et il est parfois possible d'en fixer le nombre.

Un grand nombre de méthodes de classifications non supervisées ont été appliquées aux documents textuels. Dans cet article, nous proposons la méthode des cartes auto-organisatrices de Kohonen pour la classification de documents textuels basés sur les n-grammes.

La section 2 introduira les différentes manières pour représenter un texte, expliquera les mesures de similarité et passera en revue les algorithmes de « clustering » les plus connus. Dans la section 3, nous décrivons l'approche proposée avec toutes ses étapes. Enfin la section 4 conclura l'article.

## 2 Etat de l'art

Mettre en œuvre l'une de ces méthodes consiste en premier lieu à choisir une manière de représenter les documents [1], car il n'existe pas actuellement de méthode d'apprentissage capable de traiter directement des données non structurées (textes). Dans un second temps il faut choisir une mesure de similarité, et en dernier lieu choisir un algorithme de classification non supervisée que l'on va mettre au point à partir des descripteurs et de la métrique choisis.

### 2.1 Représentation des documents Textuels

Pour mettre en œuvre n'importe quelle méthode de classification il faut d'abord transformer les textes numérisés de manière à la fois économique et significative afin qu'ils soient analysables. Le modèle vectoriel est l'approche la plus courante pour représenter les documents textuels: on représente un texte par un vecteur numérique obtenu en comptant les éléments lexicaux les plus pertinents.

Tout document  $d_j$  sera transformé en un vecteur :

$$d_j = (w_{1j}, w_{2j}, \dots, w_{|T|j}) \quad (1)$$

où  $T$  est l'ensemble de termes (ou descripteurs) qui apparaissent au moins une fois dans le corpus dit aussi taille du vocabulaire, et  $w_{kj}$  représente le poids (fréquence ou importance) du terme  $t_k$  dans le document  $d_j$ .

Table 1. Matrice documents-termes

Documents	Termes ou Descripteurs					
$d_1$	$w_{11}$	$w_{21}$	$w_{31}$	...	$w_{j1}$	...
$d_2$	$w_{12}$	$w_{22}$	$w_{32}$	...	$w_{j2}$	...
...	...	...	...	...	...	...
$d_m$	$w_{1m}$	$w_{2m}$	$w_{3m}$	...	$w_{jm}$	...

- La représentation de textes la plus simple, introduite dans le cadre du modèle vectoriel porte le nom de « sac de mots » ou « bag of words » [2] [3], elle consiste à transformer les textes en vecteurs dont chaque composante représente un mot. Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots et déstructure syntaxiquement les textes en les rendant compréhensibles pour la machine.
- Une autre représentation nommée « sac de phrases » effectue une sélection des phrases (des séquences de mots se suivant dans le texte, et non pas l'unité lexicale « phrase » telle qu'on l'entend habituellement), en privilégiant celles qui sont susceptibles de porter un sens important. Logiquement, une telle représentation doit obtenir de meilleurs résultats que ceux obtenus par la représentation « sac de mots ». Mais les expériences [4] montrent que si les qualités sémantiques sont conservées, les qualités statistiques sont largement dégradées.
- Une autre méthode pour la représentation des textes fait appel aux techniques de lemmatisation et de stemming qui consistent à rechercher les racines lexicales pour les unes [5] et à remplacer les verbes par leur forme infinitive et les noms par leur forme au singulier [6] pour les autres afin d'éviter que chaque flexion ou forme d'un mot ne soit considérée comme un descripteur différent et donc une dimension de plus.
- Une autre méthode de représentation qui présente plusieurs avantages est la méthode basée sur les « n-gram » où un « n-gram » est une séquence de  $n$  caractères consécutifs. L'ensemble des « n-gram » (généralement  $n$  varie de 2 à 5) qui peuvent être générés pour un document donné est principalement le résultat du déplacement d'une fenêtre de  $n$  caractères le long du texte [7]. La fenêtre est déplacée de un caractère à la fois, le nombre d'occurrences des différents « n-gram » est alors compté [8], [9].
- La représentation conceptuelle nommée aussi représentation basée sur une ontologie, se base aussi sur le formalisme vectoriel pour représenter les documents mais elle reste fondamentalement différente des méthodes de représentation présentées auparavant. La propriété de cette approche réside dans le fait que les éléments de l'espace vectoriel ne sont pas associés ici à des termes d'indexation seulement mais à des



concepts aussi, cela est possible en ajoutant une étape supplémentaire, c'est l'étape de mapping des termes dans les concepts.

Il existe différentes méthodes pour calculer le poids  $w_{kj}$  en sachant que, pour chaque terme, il est possible de connaître d'une part sa fréquence d'apparition dans le corpus mais également le nombre de documents qui contiennent ce terme.

La majorité des approches [1] sont axées sur une représentation vectorielle des textes de type  $TF-IDF$ .

La fréquence  $TF$  d'un terme  $t$  dans un corpus de documents textuels correspond au nombre d'occurrences du terme  $t$  dans le corpus. La fréquence  $IDF$  d'un terme  $t$  dans un corpus de documents textuels correspond au nombre de documents contenant  $t$ . Ces deux notions sont combinées multiplicativement de façon à attribuer un poids d'autant plus fort que le terme apparaît souvent dans le document et rarement dans le corpus complet.

$$TF \times IDF(t_k, d_j) = Occ(t_k, d_j) \times \log \frac{Nbre\_doc}{Nbre\_doc(t_k)} \quad (2)$$

où  $Occ(t_k, d_j)$  est le nombre d'occurrences du terme  $t_k$  dans le document  $d_j$ ,  $Nbre\_doc$  est le nombre total de documents du corpus et  $Nbre\_doc(t_k)$  est le nombre de documents de cet ensemble dans lesquels apparaît au moins une fois le terme  $t_k$ .

Il existe une autre mesure de pondération appelée  $TFC$  similaire à  $TF \times IDF$  qui en plus corrige les longueurs des textes par la normalisation en cosinus, pour éviter de favoriser les documents les plus longs.

$$TFC(t_k, d_j) = \frac{TF \times IDF(t_k, d_j)}{\sqrt{\sum_{k=1}^{|T|} (TF \times IDF(t_k, d_j))^2}} \quad (3)$$

## 2.2 Mesure de similarité

Typiquement, la similarité entre documents est estimée par une fonction calculant la distance entre les vecteurs de ces documents, ainsi deux documents proches selon cette distance seront considérés comme similaires. Plusieurs mesures de similarité ont été proposées [10]. Parmi ces mesures on peut citer :

- La distance du cosinus :

$$Cos(d_i, d_j) = \frac{\sum [TF \times IDF(t_k, d_i)] \cdot [TF \times IDF(t_k, d_j)]}{\|d_i\|^2 \cdot \|d_j\|^2} \quad (4)$$

- La distance Euclidienne :

$$Euclidian(d_i, d_j) = \sqrt{\sum_{k=1}^n (w_{ki} - w_{kj})^2} \quad (5)$$

- La distance de Manhattan :

$$Manhattan(d_i, d_j) = \sum_{k=1}^n |w_{ki} - w_{kj}| \quad (6)$$

### 2.3 Algorithmes de classification non supervisée de documents textuels

La classification non supervisée ou « clustering » est l'une des techniques fondamentales du datamining pour regrouper des données structurées ou non structurées. Plusieurs méthodes ont été proposées, selon [11] et [12], ces méthodes peuvent être classées comme suit :

▪ *Méthodes hiérarchiques :*

Ces méthodes génèrent un arbre hiérarchique de classes appelé dendrogramme. Il y a deux manières de construire l'arbre: à partir des documents ou à partir de l'ensemble de tous les documents ou corpus.

- Si on se base sur les documents, chaque document est d'abord mis dans une classe qui n'en contient qu'un seul. Ensuite, les deux classes les plus similaires sont fusionnées pour n'en former qu'une. Ce processus se répète itérativement jusqu'à ce qu'une certaine condition d'arrêt soit satisfaite. Cette méthode est appelée « agglomération de groupes similaires » ou « Clustering » hiérarchique ascendant.
- Par contre, si on se base sur l'ensemble des documents ou corpus, la méthode est appelée « division de groupes dissimilaires » ou « Clustering » hiérarchique descendant. Au début de ce processus, il n'y a qu'une classe de tous les documents. La classe est divisée en deux sous-classes lors de l'itération suivante. Le processus continue jusqu'à ce que la condition d'arrêt soit satisfaite. La similarité entre deux documents se base sur la distance entre documents.

▪ *Méthodes par partitionnement :*

Appelées aussi « clustering » à plat. Les méthodes connues sont la méthode des K-médoïds, la méthode des nuées dynamiques et la méthode des K-means ou centres mobiles.

Pour la méthode des K-means par exemple le nombre de classes est prédéfini. Un document est mis dans une classe si la distance entre le vecteur du document et le centre de cette classe est la plus petite en comparaison avec les distances entre le vecteur et les centres des autres classes.

▪ *Méthodes à base de densité :*

Il s'agit de grouper les objets tant que la densité de voisinage excède une certaine limite. Les groupes ou classes sont des régions denses séparés par des régions peu denses. Un point (vecteur de document) est dense si le nombre de ses voisins dépasse un certain seuil et un point est voisin d'un autre point s'il est à une distance inférieure à une valeur fixée.

La découverte d'un groupe ou classe se déroule en 2 étapes :

- Choisir aléatoirement un point dense
- Tous les points qui sont atteignables à partir de ce point, selon le seuil de densité, forment un groupe ou une classe

▪ *Méthodes à base de grille :*

Il s'agit de diviser l'espace de données en cellules formant une grille multidimensionnelles (les données sont représentées comme des points dans la grille) et grouper les cellules voisines en terme de distance. Les classes sont construites en assemblant les cellules contenant suffisamment de données (denses). Plusieurs niveaux de grilles sont utilisés, avec une résolution de plus en plus élevée.

▪ *Méthodes à base de modèles :*

L'une des méthodes à base de modèles est l'approche conceptuelle. Dans cette approche il s'agit d'une hiérarchie conceptuelle inhérente aux données où le concept est le couple (intention, extension) sachant que l'intention est l'ensemble maximal d'attributs communs aux vecteurs et l'extension est l'ensemble maximal de vecteurs partageant les attributs.

Une autre méthode à base de modèles est la méthode des réseaux Kohonen appelée aussi cartes auto-organisatrices de Kohonen ou encore Self Organizing Maps (SOM). C'est une méthode neuronale présentant l'intérêt d'ordonner topologiquement les classes obtenues sous forme d'une carte, généralement sur un plan (bidimensionnelle).

### 3 Approche proposée

#### 3.1 Corpus

Les données utilisées dans nos expérimentations sont issues des textes du corpus Reuters-21578 qui est un ensemble de dépêches financières émises au cours de l'année 1987 par l'agence Reuters en langue anglaise et disponible gratuitement sur le web. Ce corpus est une mise à jour effectuée en 1996 du corpus Reuters-22173. Les textes de ce corpus ont un style journalistique rédigés et sont souvent utilisés comme base de comparaison entre les différents outils de classification de documents. Sa particularité est que chaque document est étiqueté par plusieurs classes. Nous avons utilisé ces textes après avoir procédé à quelques modifications dans la phase prétraitement.

#### 3.2 Une approche basée sur les n-grammes pour la représentation des textes

Dans une première phase, on élimine des textes, la ponctuation telle que : le point, la virgule, le point virgule, le point d'interrogation et d'exclamation...etc, car ces caractères n'ont pas d'influence sur les résultats de la classification et n'apportent pas d'informations pertinentes pour la prise de décision, aussi leur élimination réduit la dimension de l'espace de représentation

On compte ensuite les fréquences des n-grammes trouvés, où un n-gram est une séquence de n caractères consécutifs. Pour un document donné, l'ensemble des n-grammes (en général  $n = \{2, 3, 4, 5\}$ ) est le résultat qu'on obtient en déplaçant une fenêtre de n cases sur le corps du texte. Ce déplacement se fait par étapes de un caractère et à chaque étape on prend une « photo » l'ensemble de ces « photos » donne l'ensemble de tous les n-grammes du document.

Par exemple, pour générer tous les 5-grams dans la phrase : « *the\_fisherman\_fish* ».

On obtient:

[ the\_f=1, he\_fi=1, e\_fis=1, \_fish=2, fishe=1, isher=1, sherm =1, herma=1, erman=1, rman\_ =1, man\_f=1, an\_fi=1, n\_fis=1 ]

Le caractère \_ est utilisé pour représenter un blanc.

Avec l'utilisation des n-grammes pour la représentation des documents textuels, on n'a pas besoin de faire un prétraitement linguistique, c'est-à-dire, on n'a pas besoin d'appliquer des techniques de lemmatisation et de stemming ou d'éliminer les «mots vides» (stop words). Cette méthode offre d'autres avantages tels que la capture des racines des mots les plus fréquents, opère indépendamment des langues et tolère les erreurs dues aux fautes d'orthographe et aux documents scannés en utilisant un OCR.

Pour calculer le poids (fréquence) de chaque n-gram, on utilise la fonction *TFIDF*. Chaque document sera ainsi représenté par son vecteur de n-grammes normalisé.

#### 3.3 Les cartes auto-organisatrices de Kohonen (SOM) pour la classification non supervisée de documents textuels

La SOM (Self Organizing Maps ou carte auto-organisatrice de Kohonen) est une méthode d'apprentissage non supervisé qui se base sur le principe de la compétition selon un processus itératif de mises à jour [13], [14], [15].

Le modèle ou réseau de Kohonen proposé par Tuevo Kohonen [16] est une grille (carte) généralement bidimensionnelle de  $p$  par  $p$  unités (cellules, nœuds ou neurones)  $N_i$ . Il est constitué:

- D'une couche d'entrée: tout objet à classer est représenté par un vecteur multidimensionnel (le vecteur d'entrée). A chaque objet est affecté un neurone qui représente le centre de la classe.
- D'une couche de sortie (ou couche de compétition). Les neurones de cette couche entrent en compétition pour être activés selon une distance choisie, un seul neurone est activé (*winner-takes-all neuron*) suite à la compétition.

L'algorithme SOM a été depuis longtemps proposé et appliqué dans le domaine de la classification de documents textuels. Les recherches concernant la SOM sont très nombreuses aujourd'hui [17].

Cependant, la combinaison entre SOM et représentation basée sur les n-grammes n'a pas été beaucoup étudiée. Dans ce qui suit, nous allons expérimenter cette combinaison et faire son évaluation.

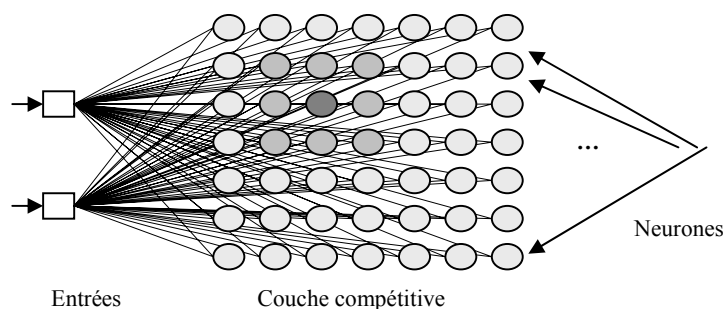


Fig. 1. Architecture du réseau de kohonen

### 3.4 Configuration

Notre système a été développé avec Borland JBuilder version7 sous la plate-forme Windows XP, sur une machine muni d'un processeur Intel Pentium 4 (2,66 GHz) avec 256 Mo de RAM.

Nous avons utilisé une carte de Kohonen 7x7 et nous avons testé pour n variant de 2 à 5 deux mesures de similarité : la distance euclidienne (distance par défaut dans l'algorithme de Kohonen) et la distance du cosinus qui est couramment employée dans les systèmes de recherche d'information et dans la classification de textes.

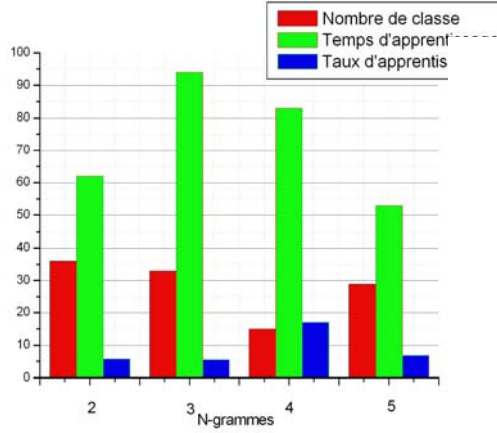
### 3.5 Résultats

Il est nécessaire de spécifier ici que notre objectif est de montrer qu'il est possible d'étendre l'utilisation des n-grammes à la classification non supervisée de textes.

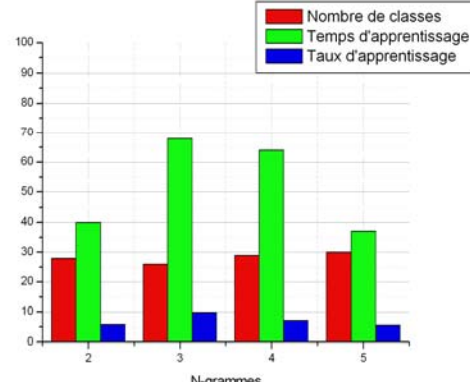
Pour chacune des valeurs de n et pour chaque mesure de similarité, citées ci-dessus, nous avons calculé le nombre de classes, le temps et le taux d'apprentissage. Nous avons obtenu les résultats suivants :

**Table 2.** Nombre de classes, Temps d'apprentissage et Taux d'apprentissage en fonction des deux mesures de similarité pour n= {2,3,4,5}

	2	3	4	5	2	3	4	5
	Distance du cosinus				Distance euclidienne			
Nombre de classes	36	33	15	29	28	26	29	30
Temps d'apprentissage (en S)	62	94	83	53	40	68	64	37
Taux Apprentissage Maximal (%)	5,81	5,62	17,05	6,95	5,81	9,71	7,14	5,61



**Fig. 2.** Nombre de classes, Temps d'apprentissage et Taux d'apprentissage en fonction de la distance du cosinus pour  $n = \{2, 3, 4, 5\}$



**Fig. 3.** Nombre de classes, Temps d'apprentissage et Taux d'apprentissage en fonction de la distance euclidienne pour  $n = \{2, 3, 4, 5\}$

On remarque que les meilleurs résultats, en termes de nombre de classes et de taux d'apprentissage, sont obtenus pour  $n=3$  avec la distance euclidienne et pour  $n=4$  avec la distance du cosinus.

### 3.6 Evaluation

L'évaluation de la pertinence des classes formées reste un problème ouvert. La difficulté vient principalement du fait que cette évaluation est subjective par nature car il existe souvent différents regroupements pertinents possibles pour un même jeu de données. Néanmoins, Il y a en général 4 critères principaux pour évaluer une classification non supervisée de documents textuels:

- La capacité de traiter de très gros volumes de données non structurées.
- L'interprétation des résultats : le système doit offrir différents modes de visualisation des résultats. Dans notre approche la carte de Kohonen est un bon exemple de visualisation.
- Chaque groupe doit être le plus homogène possible, et les groupes les plus différents possibles entre eux. Cela revient à choisir la mesure de similarité la mieux adaptée.
- Une bonne représentation influence indéniablement le clustering.

Si l'on souhaite évaluer la qualité d'une classification non supervisée en fonction des classes connues pour chaque document, deux mesures de la qualité externe sont classiquement utilisées : la f-mesure et l'entropie. Ces deux mesures sont basées sur deux notions : le rappel et la précision tels que :

$$Précision(i, k) = \frac{N_{ik}}{N_k} \quad (7)$$

$$Rappel(i, k) = \frac{N_{ik}}{N_{Ci}} \quad (8)$$

Où  $N$  est le nombre total de documents,  $i$  est le nombre de classes (prédéfinies),  $K$  est le nombre de clusters de la classification non supervisée,  $N_{Ci}$  est le nombre de documents de la classe  $i$ ,  $N_k$  est le nombre de documents du cluster  $C_k$ ,  $N_{ik}$  est le nombre de document de la classe  $i$  dans le cluster  $C_k$ .

La F-mesure  $F(P)$  et l'Entropie sont calculées comme suit :

$$F(P) = \sum \frac{N_{Ci}}{N} \max_{k=1}^K \frac{(1 + \beta) \times Rappel(i, k) \times Précision(i, k)}{\beta \times Rappel(i, k) + Précision(i, k)} \quad (9)$$

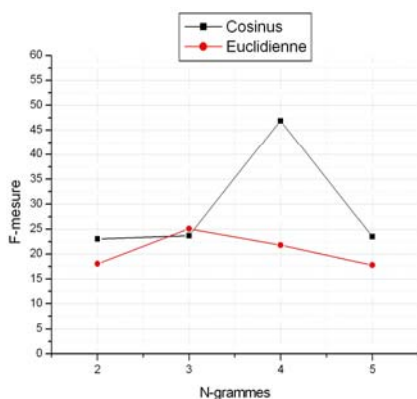
$$E(P) = \sum_{k=1}^K \frac{N_k}{N} \times \left( - \sum_i \text{Précision}(i,k) \times \log \text{Précision}(i,k) \right) \quad (10)$$

Typiquement  $\beta=1$ .

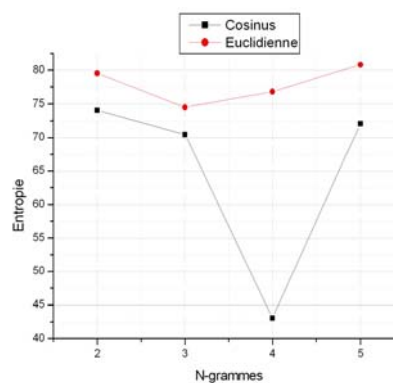
La partition P considérée comme la plus pertinente et qui correspond le mieux à la solution externe attendue est celle qui maximise la F-mesure associée ou minimise l'entropie associée.

**Table 6.** Comparaison des valeurs de la F-mesure et de l'entropie obtenues pour  $n = \{2, 3, 4, 5\}$

%	Distance	n-grammes			
		2	3	4	5
F-mesure	Cosinus	23,04	23,7	<b>46,9</b>	23,48
	Euclidienne	18,07	25,1	21,8	17,8
Entropie	Cosinus	74,12	70,49	<b>43,05</b>	72,1
	Euclidienne	79,56	74,53	76,83	80,83



**Fig. 4.** Comparaison des valeurs de la F-mesure obtenues pour  $n = \{2, 3, 4, 5\}$



**Fig. 5.** Comparaison des valeurs de l'entropie obtenues pour  $n = \{2, 3, 4, 5\}$

Nous remarquons que la meilleure performance est obtenue pour  $n=4$  avec la distance du cosinus. Nous constatons que les performances s'améliorent lorsqu'on augmente la valeur de  $n$  jusqu'à  $n=4$  et commencent à se dégrader à partir de  $n=5$ . Nous confirmons ainsi les travaux antérieurs [9], [18], qui ont montré que l'approche avec les 4-grammes produit de meilleurs résultats que celle des 3-grammes, où moins d'objets sont générés. D'autre part, les besoins en mémoire et la complexité des 5-grams font que cette approche soit généralement inacceptable.

## 4 Conclusion

Dans cet article nous avons présenté la notion de classification automatique non supervisée de textes avec ces étapes : la représentation, le choix d'une métrique et le choix de la méthode. Il est à noter que les méthodes de représentation doivent bénéficier de la même importance que les méthodes de classification cela s'explique par le fait qu'une bonne classification nécessite une bonne représentation [4]. Nous avons proposé une nouvelle approche pour la classification non supervisée de textes SOM basée sur les  $n$ -grammes. Les résultats obtenus montrent que l'approche avec les 4-grammes combinés à la distance du cosinus permet de construire une classification de meilleure qualité. Nous projetons d'utiliser la particularité des  $n$ -grammes à opérer indépendamment des langues dans la classification par la méthode SOM pour des textes multilingues.

## Références

1. Sebastiani, F.: "Machine learning in automated text categorization". *ACM Computing Surveys*, 34(1), 1–47, 2002. Available from World Wide Web: <http://faure.iei.pi.cnr.it/~fabrizio/Publications/ACMCS02.pdf>
2. Salton, G., McGill, M.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York 1983.
3. Aas, K., Eikvil, L.: "Text categorization: a survey". *Technical report, Norwegian Computing Center*, 1999. Available from World Wide Web: [http://www.nr.no/documents/samba/research\\_areas/BAMG/Publications/tm\\_survey.ps](http://www.nr.no/documents/samba/research_areas/BAMG/Publications/tm_survey.ps)
4. Schütze, H., Hull, D. A., Pedersen, J.O.: "A comparison of classifiers and document representations for the routing problem". In: *Fox, E. A., Ingwersen, P., and Fidel, R., editors, Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, Seattle, US. ACM Press, New York, 229–237, 1995. Available from World Wide Web: <ftp://parcftp.xerox.com/pub/qca/papers/sigir95.ps.gz>
5. Sahami, M.: "Using Machine Learning to Improve Information Access". *PhD thesis*, Computer Science Department, Stanford University, 1999.
6. de Loupy, C.: "L'apport de connaissances linguistiques en recherche documentaire". In: *TALN'01*, 2001.
7. Miller, E., Shen, D., Liu, J., Nicholas, C.: "Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System". *Journal of Digital Information*, 1(5), 1999.
8. Elberichi, Z.: "Text mining using n-grams". In: *Proceedings of CIIA'06*, Saida Algeria, May 2006.
9. Rahmoun, A., Elberichi, Z.: Experimenting N-Grams in Text Categorization. *Issue of International Arab Journal of Information Technology*, Vol 4, N° 4, pp. 377-385, 2007.
10. Jones, W. and Furnas, G.: "Pictures of relevance: A geometric analysis of similarity measures". *Journal of the American Society for Information Science*, 38(6):420-442, November 1987.
11. Berkhin, P.: *Survey of Clustering Data Mining Techniques*. Accrue Software, San Jose CA, 2002.
12. Wang, Y.: "Incorporating semantic and syntactic information into document representation for document clustering". *A dissertation submitted to the Faculty of Mississippi State University August, 2002*.
13. Amine, A., Elberichi, Z., Simonet, M., Malki, M.: "Classification non supervisée de documents textuels : état de l'art". In : *Proceedings of COSI'07*, Oran Algeria, June 2007.
14. Amine, A., Elberichi, Z., Simonet, M., Malki, M.: SOM-based Clustering of Textual Documents Using WordNet. *International Arab Conference on Information Technology, Science & Tech. and Maritime Trans.*, Lattakia, Syria, 2007.
15. Amine, A., Elberichi, Z., Simonet, M., Malki, M.: SOM pour la Classification Automatique Non supervisée de Documents Textuels basés sur Wordnet. EGC2008, INRIA Sophia Antipolis, France.
16. Kohonen, T.: *Self-organized formation of topologically correct feature maps*. *Biological Cybernetics*, 43, 59-69, 1982.
17. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A.: *Self organization of a massive document collection*. *IEEE Transactions on Neural Networks*. 11(3), May 2000.
18. Bayer, Th., H. Mogg-Schneider, I. Renz, H. Schäfer, 1997. Daimler Benz Research: System and Experiments Routing and Filtering. In *Proceedings of the 6th Text REtrieval Conference TREC-97*.

## Using Word-Space Models for Enriching Multilingual Lexical Resources and Detecting the Relation Between Morphological and Semantic Composition

Adil Toumouh<sup>1</sup>, Dominic Widdows<sup>2</sup>, Ahmed Lehireche<sup>1</sup>

<sup>1</sup> Computer Science Departement, University of Djillali Liabes, Sidi Bel Abbès, Algeria  
toumouh@gmail.com

<sup>2</sup> Google, Inc.  
widdows@google.com

**Abstract.** Over the past 15 years, Word-Space models have been increasingly used as practical ways of capturing and representing distributional information about words. Indeed, various experiments reported in the literature demonstrate that Word-Space models can be used to create or enrich Ontologies. However, the results of such experiments have been difficult to formalize, or to demonstrate that Word-Space based methods are accurate enough to reliably replace or at least supplement the work of professional knowledge engineers. In this paper we review two specific proposals, in which we hope to demonstrate that Word-Space models can be used for clearly defined linguistic tasks such as learning translation pairs and producing meaning representations for morphological compositions. As motivation, we describe the family of Word-Space models, review experiments in the literature that give us initial confidence, and explain how these can be taken much further.

**Keywords:** Word-Space Model, Random Indexing, LSA, Parallel Corpora, Multilingual resources, Ontology, Bilingual vector Model, Translation, Morphological composition.

### 1 Introduction: Word-Space model

A Word-Space model is an algebraic model which provides a geometric representation of words in a high dimensional vector space. This geometric representation allows us to tell if two words have similar meaning by looking at their spatial proximity. In Word Space, semantically related words are close, unrelated words are distant [29]. This idea is based on the distributional hypothesis: *Words with similar distributional properties have similar meanings*. This does not mean that the words occur with each other, but only requires that the words co-occur with a similar pattern of distributional neighbors words [27].

The multi-dimensional space is constructed only by using the distributional statistic extracted from large collections of text documents. In economic terms, the Word-Space model therefore compares favorably with other models of meaning, because the model is constructed with no human intervention. The only input data needed to build a Word-Space model is significant amounts of free “unstructured” text – and as recent years have shown, this is one resource that we have in abundance for an increasing number of languages. Word-Space models are building with no a priori knowledge or constraints about meaning similarities, there are no predefined classes or types of words. While this makes for scalable and robust methods, it also presents us with the significant challenge of integrating information from a Word-Space model with information contained in more formalized models which do have predefined notions of classes, types and relations (e.g. WordNet [9]).

Methods for building Word-Space models usually have two phases:

- i. Gathering co-occurrence data for each word with a given set of features.
- ii. Reducing dimensions from the number of features (which is usually very large) to a smaller number in a “reduced space”.

In the following sections 2 and 3, we outline some of the main options for accomplishing each of these phases.



## 2 Gathering Co-occurrence Data and Dimension reduction

The most standard form of co-occurrence data in this field is the traditional *term-document matrix* used in information retrieval [3] [28]. For a large corpus of documents, a term-document matrix records the number of times each term occurred in each document. We can think of this either as describing each document as a collection of terms, so that the documents are the objects and the terms are the features. However, we can also think of this as describing a *term* as a collection of *documents* – in which case, the terms are the objects and the documents are the features. The dimensionality of the Word-Space is determined by the number of columns. Each row is called a context vector. Computing similarity between words become as computing similarity between their vectors representations by using a kind of similarity measures, such as the cosine similarity.

Unfortunately, the size of the documents we manipulate are very large, which make the dimensionality of the matrix very high. Consequently, the matrix will be intractable and affects the scalability of this method. An ideal solution is to preserve a big size of data and to project them in lower dimensional space. However, using reduction technique of dimensionality will affect the integrity of document information's. For this reason, research undertaken focused on reducing the dimensionality efficiently and preserving as much information from document. These dimension reduction techniques are applied as a data pre-processing step. Several techniques were born, among which have quotes "latent semantic analysis" LSA [16] and "Hyperspace Analog to Language" HAL [18]. LSA have enjoyed a large popularity and has been so attractive for numerous fields of research, for instance information retrieval [10]. LSA performs reduction of space dimension by using SVD Singular Value Decomposition. It is a technique based on factorizing the initial matrix into three smaller matrices. The basic steps of LSA are:

1-Building a words-by-documents matrix: this process was explained in previous section.

2-Singular Value Decomposition (SVD): it is a well-known dimensionality reduction technique. The initial matrix is decomposed in three matrices:

$$\mathbf{X}_{[m \times n]} = \mathbf{L}_{[m \times r]} * \mathbf{S}_{[r \times r]} * \mathbf{R}_{[r \times n]}$$

$\mathbf{L}$  and  $\mathbf{R}$  are orthogonal matrices, whose columns are called left singular vector and right singular vector respectively.  $\mathbf{S}$  is a diagonal matrix with singular values on the diagonal.

3-Dimension Reduction: If the singular values are sorted in descending order, the task of reducing is performed by removing dimensions corresponding to small singular values and keeping those of larger singular values. The new decomposition becomes :

$$\mathbf{X}'_{[m \times n]} = \mathbf{L}'_{[m \times k]} * \mathbf{S}'_{[k \times k]} * \mathbf{R}'_{[k \times n]}$$

The resulting matrix  $\mathbf{X}'$  is an approximation of  $\mathbf{X}$ , and the space dimension becomes  $k$ .

## 3 Random Indexing

The solution of reducing the matrix dimension remains a costly operation and can't be applied before constructing the huge matrix. If we need to add new data, we are obliged to repeat the entire building process and to reduce the matrix from the beginning.

Random indexing overcomes these scalability and efficiency problems by incrementally accumulating  $k$ -dimensional index vectors into a context matrix of order  $w \times k$ , where  $k \ll d$  (the order of the original matrix is  $w \times d$ ) [25].

The random indexing methodology process as follow:

Each context is assigned a unique and random index vector of ternary values (0, +1, -1). These vectors consist of a large number of 0 and a small number (about 1-2%) of +1s and -1s randomly distributed. The dimensionality of these randomly generated vectors is usually chosen to be on the order of thousands.

As the corpus is scanned, the context vector for each word is obtained by summing the index vectors of all the contexts in which the word appears.

For instance: suppose that we have three documents D1, D2, D3, each one is assigned respectively the three random index vectors: {00+100...} {-10000} {0+100-1...}. If a word  $w$  occurs in the three documents, the context vector of this word will be the sum of the three vectors: {-1+1+10-1...}

The underlying idea of this technique and others, such as Random Projection [23] and Random Mapping [14] is the Johnson-Lindenstrauss lemma [13]: if we project points in a vector space into a randomly selected subspace of sufficiently high dimensionality, the distances between the points are approximately preserved.

Random Indexing has many advantages compared to other methods [26]:

- it is an incremental method, which means that we do not have to sample all the data before we can start using the context vectors,
- it avoids the “huge matrix step”, since the dimensionality  $k$  of the vectors is much smaller,
- it is scalable, since adding new contexts to the data set does not increase the dimensionality of the context vectors.

## 4 Parallel corpora and cross lingual context

A parallel corpus is a collection of documents translated into more than one language and it represents a rich source for machine translation and multilingual natural language processing. Many of the most useful parallel corpora have arisen for political reasons: for example, the proceedings of the Canadian parliament are recorded in both English and French, and versions of many official European Union documents are produced in not just two but several languages [35]. They represent resources for automatic lexical acquisition (e.g., [10] and [19]), they provide indispensable training data for statistical translation models (e.g., [5] [20] [22]) and they can provide the connection between vocabularies in cross-language information retrieval (e.g., [15] [21] [24]).

Other field, for which the parallel corpora play a key source, is Machine Translation. The beginning of modern Statistical Machine Translation (SMT) can be traced back to 1988, when Brown et al. [4] from IBM published a formalized mathematical formulation of the translation problem and proposed five word alignment models – IBM models 1, 2, 3, 4 and 5. Starting with a bilingual parallel sentence-aligned corpus, the IBM models learn how to translate individual words and the probabilities of these translations.

As modern Machine Translation System we cite Google's Statistical Machine Translation project. The Statistical Machine Translation project is a research project at Google which has the goal to significantly improve the quality of machine translation beyond the existing state-of-the-art while at the same time making it easier to develop machine translation systems for new language pairs. The goal is to build machine translation systems based on statistical techniques which learn from parallel corpora.

## 5 Building a Bilingual Vector Model from Parallel Corpus

We have seen in the section 2, the standard process whereby a Word-Space model can be built from monolingual documents. Our work is based on parallel corpora. We are interested by the bilingual vector model developed by Widdows et al. [33]. Before introducing their model, we demonstrate that building a bilingual vector model can be viewed as the operation of mapping between two different Word Spaces.

Suppose we have two different Word-Space models  $S$  and  $T$ , and we wish to map the vector  $\mathbf{v}$  from  $S$  to an appropriate vector in  $T$ . Now, we suppose that the neighbors of  $\mathbf{v}$  in the source model  $S$  include the vectors  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$ , and that these words also have vectors  $\mathbf{a}'$ ,  $\mathbf{b}'$  and  $\mathbf{c}'$  in the target model  $T$ . We assume the distances  $d(\mathbf{v}, \mathbf{a})$ ,  $d(\mathbf{v}, \mathbf{b})$  and  $d(\mathbf{v}, \mathbf{c})$  in the source model  $S$ : we then try to find the vector  $\mathbf{v}'$  in the target model  $T$  so that the new distances in  $d(\mathbf{v}', \mathbf{a}')$ ,  $d(\mathbf{v}', \mathbf{b}')$  and  $d(\mathbf{v}', \mathbf{c}')$  match the old distances  $d(\mathbf{v}, \mathbf{a})$ ,  $d(\mathbf{v}, \mathbf{b})$  and  $d(\mathbf{v}, \mathbf{c})$  as closely as possible.

The main idea in building a bilingual WordSpace is to consider each document along with its translation as if they were a single ‘compound document’. In [33] each German/English document pair is treated as a single ‘compound document’ for the purpose of recording term-term co-occurrence. Two words that occur in documents which are translations of one another can be regarded as co-occurring, just as if they had occurred in the same monolingual document. The 1000 most frequent English words have been selected as content-bearing words after removing all stop words [3]. English and German words were regarded as co-occurring with a particular content-bearing word if they occurred in the same document as the content bearing word, or

the translation of this document. The 10,000 most frequent words in each language were mapped on a single 1,000 dimensional vector. SVD was used to reduce the number of dimensions to 100. The semantic similarity was computed between English and German terms using cosine similarity [33]. The model we propose, is similar to the one describe above, but is distinguished to be based on random indexing rather than SVD.

## **6 Enriching Multilingual Resources**

Lexical resources and Ontologies are time consuming and labor-intensive to maintain. For this reason, during recent years there has been an increased interest to acquire or extend lexical resources and ontologies [11] [7] [12] [17] [31] [6] [1] [2] [8]. The goal of automatically placing new words into taxonomy has been attempted in various ways. Many approaches have been developped for enriching monolingual resources [34], [9] and [12]. Unlikely the previous approaches for enhancing a monolingual resource, Widdows et al. [33] proposed an approach for enhancing a multilingual resource. The bilingual vector model generates candidate German translations of English words using the cosine similarity measure between terms in the bilingual vector space. The missing German translations of English words are added to the Unified Medical Language System (UMLS), a publicly available medical language resource. The results that obtained show that accuracy of translation for high-scoring pairs can exceed 90%, and that the method also finds synonyms of terms which are already contained in UMLS [33]. The multilingual resources represent a rich mine of knowledge, and deserves to have more intention. We aim to follow the same method developed by [33], but this time by using another technique of reduction which is random indexing.

## **7 Morphological and semantic composition**

Schone and Jurafsky [30], demonstrate that some of the encountered problems – for instance applying inappropriately valid affixes: “ally” stemming to “all”- in the previous morphology induction approaches can be resolved by incorporating word semantics, and thereafter, “all” is not semantically similar to “ally”. [30] Shows in their approaches that incorporating LSA-based semantics alone into the morphology-induction process can provide results that rival a state-of-art system based on stem-and-affix statistics [30].

In the same context, we aim to combine Semantics and Morphology for comparing meaning representations. Morphologically related words (typically) have related meanings. However, it is far from obvious how to use such relationships in practice. For example, in spite of having a simple singular/plural relationship, "hostility" is used to mean ill-will, whereas "hostilities" means acts of warfare.

One of the prime uses for morphological tools in semantic applications has been for "stemming" in information retrieval. If words with the same stem are treated as an equivalence class, documents containing this stem will be found which would be missed by a simple string-matching search strategy. Unfortunately, stemming for English has often been shown to do as much harm as good, due to frequent "semantics/morphology mismatches" such as the one described above.

The situation with other languages is markedly different. For example, several experiments with German have demonstrated that the use of compositional morphology (often referred to as "decompounding") is invaluable in correctly representing the meanings of terms and documents.

The purpose of our project would be to investigate the relationship between morphological and semantic composition, specifically comparing English and German terms. We hope to fulfill two goals.

Firstly, the task of building automatic translation lexicons is of widespread interest and usefulness. We have recently demonstrated that a vector-semantic approach based upon translated documents performs well at extracting German-English pairs of words with the same meanings. However, the mapping from English words to German words is frequently many-to-one. To an extent, the vector-translation system predicts this (e.g. English "middle" + "ear" -> German "mittelohr", etc.).

Our project will therefore investigate the relationship between compositional morphology and such many-to-one maps. A broader reaching goal is to use these ideas to develop a more general approach to semantic composition. This is regarded as something of a "Holy Grail" for corpus-derived meaning representation - it is

much easier to represent the meanings of individual words than to put these meanings together into a sentence. (This is why commercial search engines are still based upon string-matching and "bag of words" models.) We propose to investigate models for composition in a rigorous fashion, by using known German compound terms to evaluate the accuracy of English semantic composition.

## Conclusion

Word Space technologies have already been shown to contribute to ontology learning in English, and some initial experiments have demonstrated their potential for work with other languages. Since we cannot expect the human labor used with English resources to be repeated for all other important languages, the need for automatic methods to construct resources for these languages is, if anything, even greater than for English. We believe that by careful use of resources such as parallel corpora, and linguistically motivated methods such as morphological analysis, Word Space technologies could play a major role in the development of multilingual ontologies and language technologies.

## References

1. Agirre, E., Rigau, G.: Word sense disambiguation using conceptual density. In Proceedings of COLING' 96, (1996.) pages 16–22, Copenhagen, Denmark.
2. Alfonseca, E., Manandhar, S.: Improving an ontology refinement method with hyponymy patterns. In Third International Conference on Language Resources and Evaluation, (2001) pages 235– 239, Las Palmas, Spain
3. Baeza-Yates, R., Ribiero-Neto, B.: 1999. Modern Information Retrieval. Addison Wesley / ACM Press
4. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P.: A statistical approach to language translation. In Proceedings of the 12th Annual Conference on Computational Linguistics (COLING 88). International Committee on Computational Linguistics (1988)
5. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Roossin, P.: A statistical approach to machine translation. Computational Linguistics, (1990) 16(2):79–85
6. Carballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text, in Proceedings of the 37<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, (1999) pp. 120-126
7. Dorr B., Jones D.: Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision. SIGLEX Workshop: "Breadth and Depth of Semantic Lexicons", (1996) pp. 42-50, Santa Cruz, USA
8. Faatz, A., Steinmetz, R.: Ontology Enrichment Evaluation. EKAW (2004): 497-498
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
10. Gale, W. A., Kenneth W. C.: Identifying word correspondences in parallel texts. In Fourth DARPA Workshop on Speech and Natural Language, (1991) pages 152–157, Asilomar, CA, February
11. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Norwell, MA, USA: Kluwer Academic Publishers (1994)
12. Hearst, M., Schütze, H.: Customizing a Lexicon to Better Suit a Computational Task. Proc. ACL SIGLEX Work. Acquisition of Lexical Knowledge from Text (1993)
13. Johnson, W. B., Lindenstrauss, d. J.: Extensions of Lipshitz Mapping into Hilbert Space. In Conference in modern analysis and probability, volume 26 of Contemporary Mathematics, pages 189-206. Amer. Math. Soc., (1984)
14. Kaski, S.: Dimensionality reduction by random mapping: Fast similarity computation for clustering. In Proceedings of the IJCNN'98, International Joint Conference on Neural Networks. IEEE Service Center (1999)
15. Landauer, T., Littman, M.: Fully automatic cross-language document retrieval using latent semantic indexing. In Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, (1990) pages 31–38, Waterloo, Ontario, October
16. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review (1997) 104 (2), 211-240
17. Lin, D.: Automatic Retrieval and Clustering of Similar Words. COLING-ACL98, Montreal, Canada (1998)
18. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instrumentation, and Computers, (1996) 28, 203- 208

19. Melamed, D.: Automatic discovery of non-compositional compounds in parallel data. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), (1997) pages 97–108, Providence, RI, August
20. Melamed, D.: Models of translational equivalence among words. *Computational Linguistics*, (2000) 26(2):221–249
21. Oard, D.: Cross-language text retrieval research in the USA. In Third DELOS Workshop on Cross-Language Information Retrieval, Zurich, March. European Research Consortium for Informatics and Mathematics, Sophia Antipolis, France (1997)
22. Och, F., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), (2002) pages 295–302, Philadelphia, July
23. Papadimitriou, C. H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: A probabilistic analysis. In Proceedings of the 17th ACM Symposium on the Principles of Database Systems. ACM Press (1998)
24. Resnik, R., Smith, N.: The Web as a parallel corpus. *Comput. Linguist.*, Vol. 29, No. 3. (2003), pp. 349-380.
25. Sahlgren, M.: Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data. In Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04 (2004) (pp. 1289-1292)
26. Sahlgren, M., Coster, R.: Using bag-of-concepts to improve the performance of support vector machines in text categorization. In Proceedings of the 20th International Conference on Computational Linguistics, COLING'04 (pp. 487{493) (2004)
27. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. Dissertation, Department of Linguistics, Stockholm University (2006)
28. Salton, G., McGill, M. (ed.): Introduction to modern information retrieval. McGraw-Hill (1983)
29. Schutze, H.: Word space. In Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93 (pp. 895{902). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc (1993)
30. Schone, P., Jurafsky, D.: Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop, Lisbon, (2000)
31. Takunaga T., Fujii A., Iwayama M., Sakurai N., Tanaka H.: Extending a Thesaurus by Classifying Words. "Automatic Information Extraction and Building of Lexical Semantic Resources" Workshop, Vossen P. (et al.) (eds), (1997) pp. 16- 21, Spain
32. Toumouh, A., Lehireche, A., Widdows, D., Malki, M.: Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus. Appeared in 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06) Dubai/Sharjah, UAE, March 8-11, (2006)
33. Widdows, D., B. Dorow, C.-K. Chan.: Using parallel corpora to enrich multilingual lexical resources. In Third International Conference on Language Resources and Evaluation (LREC-02), (2002) pages 240–245. Las Palmas, Spain.
34. Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In Proceedings of Human Language Technology / North American Chapter of the Association for Computational Linguistics (HLT-NAACL-03). Edmonton, Canada (2003)
35. Widdows, D.: Geometry and Meaning. CSLI Publications (2004)
36. Widdows, D., Toumouh, A., Dorow, D., Lehireche, A.: Ongoing Developments in Automatically Adapting Lexical Resources to the Biomedical Domain. Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy, May 24-26, (2006).

## *Software engineering II*

---



## Minimisation des Désagréments dans les Clusters Agrégés

Baroudi Rouba<sup>1</sup>, Safia Nait Bahloul<sup>1</sup>, Youssef Amghar<sup>2</sup>

<sup>1</sup>Département Informatique, Faculté des Sciences,  
Université Es-Sénia, Oran, Algérie

<sup>2</sup>INSA de Lyon – LIRIS FRE 2672 CNRS,  
7 avenue Jean Capelle.  
69621, Villeurbanne – France.  
[nait1@yahoo.fr](mailto:nait1@yahoo.fr)

**Résumé :** Plusieurs expériences ont montré l'impact du choix des parties de documents sélectionnées sur le résultat de la classification et par conséquent sur le nombre de requêtes auxquelles peut répondre ces classes. Le processus d'agrégation fournit une méthode très naturelle de la classification des données et considère alors les  $m$  classifications produites par les  $m$  attributs et essaie de produire une classification dite "optimale" qui est la plus proche possible des  $m$  classifications. L'optimisation consiste à minimiser le nombre de paires d'objets  $(u, v)$  telle qu'une classification  $C$  les placent dans le même cluster tandis qu'une autre classification  $C'$  les placent dans des clusters différents. Ce nombre correspond au concept de désagréments. Nous proposons une approche qui exploite les différents éléments d'un document XML participant à différentes vues pour donner des classifications différentes. Ces classifications sont ensuite agrégées en une seule minimisant le nombre de désagréments. Notre approche est divisée en deux étapes: la première consiste à appliquer l'algorithme K-means sur la collection des documents XML en considérant à chaque fois un élément différent du document. La deuxième étape agrège les différentes classifications obtenues précédemment pour produire celle qui minimise le nombre de désagréments.

**Mots-clés :** XML, classification, agrégation, désagrément.

### 1. Introduction

XML est devenu un standard pour la représentation et l'échange de données sur le web. Le nombre de documents XML échangés sur Internet augmente de plus en plus, et la quantité d'information accessible aujourd'hui est telle que les outils, même sophistiqués, utilisés pour rechercher l'information dans les documents ne suffisent plus. D'autres outils permettant de synthétiser ou classer de large collection de documents sont devenus indispensables.

La classification automatique non supervisée (ou clustering) vise à regrouper les documents similaires. Ceci permet d'accélérer le processus de recherche d'information (Information Retrieval : IR), car la recherche d'une information pertinente dans une grande collection revient alors à interroger des ensembles (classes) de taille réduite. Ceci se base sur l'idée que si un document est pertinent à une requête, ses voisinages (les documents similaires de la même classe) ont plus de chance d'être aussi pertinents.

Plusieurs expériences de classification de documents XML de structure homogène ont été réalisées par [1]. Ces expériences ont montré l'impact du choix des parties de documents sélectionnées sur le résultat de la classification et par conséquent sur le nombre de requêtes auxquelles peut répondre ces clusters. Donc l'agrégation de ces classifications permet sans aucun doute d'obtenir des clusters plus pertinents. Dans ce cadre nous proposons une approche permettant d'optimiser les clusters agrégés en minimisant le nombre de désagréments issus d'un processus de classification basé sur un ensemble d'attributs jugés pertinents.

Dans la section 2, le papier introduit la classification des documents XML et la section 3 fait la description de notre Approche, donne un certain nombre de définitions et un exemple illustratif. Dans la section 4, nous concevons l'algorithme qui résume les différentes étapes de notre approche et définissons sa complexité dans la section 5 et la section 6 conclut ce papier.

## 2. La classification des documents XML

La classification, consiste à analyser des données et à les affecter, en fonction de leurs caractéristiques ou attributs, à telle ou telle classe. Il existe une quantité importante de méthodes de classification de documents. Ces méthodes peuvent être classées généralement selon leurs objectifs en deux types :

- la classification supervisée (classement) ;
- la classification non supervisée (clustering).

On peut trouver une bonne documentation dans le livre de Tom Mitchell « Machine Learning » [2]. Les différentes présentations des méthodes de clustering sont dues, d'une part, au fait que les classes d'algorithmes se recouvrent (certaines méthodes s'appuyant, par exemple, sur des modèles probabilistes proposent un partitionnement) et d'autre part, diffèrent selon que l'on s'intéresse plutôt aux résultats du clustering (hiérarchie vs. partitionnement, clustering dur vs. clustering flou etc.), ou à la méthode utilisée pour parvenir à ce résultat (utilisation de fonctions probabilistes versus utilisation de graphes, etc...).

Plusieurs travaux concernant le clustering [3], [4], [5], [6], [7] et la similarité [8], [9], [10] des documents XML ont été réalisés, et ceci avec divers objectifs. Certains visent à identifier la partie de la DTD la plus utilisée [11], d'autres essayent d'identifier des structures fréquentes dans une large collection [12]. Parmi les objectifs, on trouve aussi le besoin d'identification des DTD pour des collections hétérogènes [13], et enfin la combinaison de la structure et du contenu des documents afin de réaliser le clustering [14]. Certaines méthodes de classification réduisent les documents XML à leur partie purement textuelle [4], [15], sans prendre avantage de la structure qui pourtant véhicule une information riche. Dans [1] ont été réalisées plusieurs expériences de classification de documents XML de structure homogène, en vue d'expliquer et de valider une présentation organisationnelle préexistante (les rapports d'activités de l'INRIA) et, l'intérêt a porté sur l'impact du choix des parties de documents sélectionnées sur le résultat de la classification. Deux niveaux de sélection ont été appliqués: une sélection utilisant la structure du document, puis une sélection linguistique au niveau du texte précédemment sélectionné. Un algorithme de classification de type k-means [16], [17] construit alors une partition des documents, affecte les documents à des classes et exhibe la liste des mots qui ont permis la classification. Ainsi ces auteurs ont montré que la qualité de la classification dépend fortement des parties de documents sélectionnées.

Plusieurs approches utilisent le concept d'agrégation en classification dans différents domaines tels que: l'apprentissage automatique [18], [19], la reconnaissance des formes [20], la bioinformatique [21], et le data mining [22], [23]. L'agrégation fournit une méthode très naturelle pour la classification des données.

En considérant un ensemble de tuples  $T_1, \dots, T_n$  caractérisé par un ensemble d'attributs  $A_1, \dots, A_m$ , notre idée consiste à voir chaque attribut comme un élément pouvant produire une classification simple de l'ensemble des données; si l'attribut  $A_j$  contient  $K_j$  valeurs distinctes alors  $A_j$  partitionne les données en  $K_j$  clusters. Le processus d'agrégation considère alors les  $m$  classifications produites par les  $m$  attributs et essaie de produire une classification dite "optimale" qui est la plus proche possible des  $m$  classifications, c'est-à-dire minimisant le nombre de paires d'objets  $(u, v)$  telle qu'une classification  $C$  le place dans le même cluster tandis qu'une autre classification  $C_0$  les placent dans des clusters différents. Ce nombre correspond au concept de désagréments [24]. Notre approche consiste alors à agréger un ensemble de classifications basées chacune sur un attribut pertinent extrait de la DTD des documents à classifier. Chaque classification est issue de l'application de l'algorithme k-means [16], [17]. La qualité des clusters obtenus est assurée d'une part par l'efficacité de l'algorithme k-means, en tant que référence des algorithmes de classification, et d'autre part par l'optimisation (minimisation des désagréments) assurée par le concept d'agrégation.

Les sections suivantes décrivent en détail les étapes et les concepts de notre approche.

## 3. Description de notre Approche

L'approche proposée suit quatre étapes:

**Etape 1 :** Détermination de l'ensemble des éléments pertinents (la pertinence de l'élément est déterminée par sa fréquence d'apparition dans les requêtes)



**Étape 2 :** Recensement pour chaque attribut de l'ensemble des mots représentatifs (valeurs possibles de l'attribut).

**Étape 3 :** Application de l'algorithme k-means pour chaque attribut extrait dans la première étape

**Étape 4 :** Agrégation des résultats obtenus à l'étape 3.

### 3.1 Exemple illustratif

Nous illustrons notre approche à travers un exemple. Soit une collection de documents XML basée sur la DTD suivante:

<ELEMENT **Film** (Titre, Genre, Réalisateur, Acteur, Actrice, Année, Budget, Editeur)>

- La première étape dans notre approche est d'identifier les parties de la DTD les plus importants pouvant produire des clusters pertinents. Il est évident que le Titre et le budget ne constituent pas des éléments de classification. Par contre, les films peuvent être regroupés en classes selon leur genre, leur réalisateur, leur acteurs ou leur éditeur.
- L'étape suivante dans le processus est le choix des mots représentatifs de chaque attribut. Le résultat de cette étape peut avoir la forme du tableau suivant:

**Table 1.** exemple de mots représentatifs des attributs

Attribut	Mots représentatifs
Genre	Comédie, action, horreur...
Acteur	Tom Cruz, Kevin Kosner...
Réalisateur	Spielberg, Newman...
Editeur	20 Century, 3 stars...

- La troisième étape consiste à appliquer l'algorithme K-means [16, 17] sur la collection en considérant à chaque fois un attribut différent. On aura par exemple des clusters "Comédie", "Horreur", "Action" pour l'attribut Genre.
- La dernière étape est la phase d'agrégation qui permet d'agréger les classifications obtenues lors de la troisième étape (première partie). Cette étape permet de construire des clusters de type « Tous les films d'**Action** réalisés par **Spielberg** et édités par **20Century** dans lesquels a joué **Tom Cruz** »

### 3.2 Définitions

#### 3.2.1. Agrégation

Soit  $C_L = \{C_1, \dots, C_m\}$  un ensemble de  $m$  classifications. Le concept d'agrégation consiste à produire une classification  $C$  qui réalise un compromis avec les  $m$  classifications.

Les auteurs de [24] définissent l'agrégation comme un problème d'optimisation visant à minimiser le nombre de désagréments entre les  $m$  classifications.

#### 3.2.2. Désagrément

Soit  $C$  et  $C_0$  deux classifications, le désagrément est défini comme étant une paire d'objets  $(u, v)$  telle que  $C$  les place dans le même cluster tandis que  $C_0$  les place dans des clusters différents. Si  $d(C_0, C)$  est le nombre de désagréments entre  $C$  et  $C_0$ , l'agrégation consistera donc à trouver une classification  $C$  qui minimise :

$$\sum_{i=1}^m d(C_i, C) \quad (1)$$

L'équation (1) permet de calculer la distance entre une classification  $C$  et l'ensemble des classifications. Cette distance représente en fait le nombre de couple d'objets  $(V_i, V_j)$  sur lesquels les deux classifications sont en désaccord.

### Exemple

Soient  $C_1$ ,  $C_2$  et  $C_3$  des classifications.  $V_1, \dots, V_6$  sont les objets à classifier. La valeur  $K$  en entrée  $(V_i, C_j)$  exprime l'appartenance de l'objet  $V_i$  au cluster  $C_j$ . La colonne  $C$  correspond à la classification optimale qui minimise le nombre de désagréments entre les classifications  $C_1, C_2, C_3$ .

**Table 2.** exemple d'une classification optimale

	$C_1$	$C_2$	$C_3$	$C$
$V_1$	1	1	1	<b>1</b>
$V_2$	1	2	2	<b>2</b>
$V_3$	2	1	1	<b>1</b>
$V_4$	2	2	2	<b>2</b>
$V_5$	3	3	3	<b>3</b>
$V_6$	3	4	3	<b>3</b>

Dans cet exemple le nombre total de désagréments est 5: un avec la classification  $C_2$  pour le couple  $(v_5; v_6)$ , et quatre avec la classification  $C_1$  pour les couples  $(v_1; v_2)$ ;  $(v_1; v_3)$ ;  $(v_2; v_4)$ ;  $(v_3; v_4)$ . Il n'est pas difficile de déterminer la classification qui minimise le nombre de désagréments qui correspond dans cet exemple à la classification  $C_3$ .

La détermination de la classification peut donc être définie comme un problème d'optimisation visant à minimiser le nombre de désagréments.

Nous matérialisons notre approche par l'algorithme **Clust-Agregat** que nous décrivons dans la section suivante.

## 4. Algorithme Clust-Agregat

Dans ce qui suit, nous présentons un algorithme qui résume les différentes étapes de notre approche. L'algorithme accepte comme entrée un ensemble  $V$  d'objets. Chaque objet est caractérisé par un ensemble d'attributs  $A$ . L'algorithme construit un ensemble  $C$  de classifications au fur et à mesure en prenant en compte à chaque fois un attribut différent.

Algorithme : **Clust-Agregat**

Entrée:  $V$  /\*l'ensemble des objets à classifier.\*/

Sortie:  $C_f$  /\* un ensemble de clusters" la classification optimale" \*/

$A = \{A_1, \dots, A_m\}$ , /\*un ensemble d'attributs.\*/

**Début**

$C := \emptyset$ ; /\*l'ensemble des classifications à optimiser\*/

**Pour**  $i$  de 1 à  $m$  faire

$C_i := K\text{-means}/A_i$ ; /\* Appliquer K-means en considérant l'attribut  $A_i$  \*/

$C := C \cup C_i$ ;

**FinPour**

$C_f := \text{Agréger}(C, V)$ ;

**Fin**

Fonction Agréger( $C, V$ ) /\*retourne une Classification  
u, v : deux objets de  $V$ .\*/

**Début**

```

Pour i de 1 à m-1 faire
  Pour j de i+1 à m faire
     $D_v(C_i, C_j) := 0;$ 
    Pour chaque  $(u, v) \in V^2$ 
       $d_{u,v}(C_i, C_j) = \begin{cases} 1 & \text{si } C_i(u) = C_i(v) \text{ et } C_j(u) \neq C_j(v) \\ & \text{ou } C_i(u) \neq C_i(v) \text{ et } C_j(u) = C_j(v) \\ 0 & \text{sinon} \end{cases}$ 
     $d_v(C_i, C_j) := d_v(C_i, C_j) + d_{u,v}(C_i, C_j);$ 
/*distance entre les deux classification  $C_i, C_j$ */
  fin pour
fin pour

 $D(C) = \sum_{i=1}^m d_V(C_i, C)$ 
Cf := Min (D(C));
Retourner (Cf);
Fin

```

La fonction **Agréger** retourne une classification **optimale**. Le critère d'optimalité du résultat correspond à la minimisation du nombre de désagrément; en d'autre terme cette fonction retourne la classification qui est en accord avec toute les classifications de l'ensemble C.

## 5. Complexité de l'algorithme Clust-Agregat

*Complexité Globale de l'algorithme Clust-Agregat* : La complexité de Clust-Agregat dépend d'une part de la complexité de l'algorithme k-means ainsi que celle de la fonction Agréger. La complexité d'une variante de l'algorithme k-means (k-médiode flous) a été évaluée à  $O(n^2)$  au pire des cas [25]. D'autre part, le processus d'agrégation est de nature NP-complet [24], mais il a été démontré ensuite qu'il est facile de trouver 2-approximation et par conséquent réduire la complexité du processus d'agrégation à  $O(mn)$  (m le nombre de classification à agréger) (voir les détails de l'algorithme BESTCLUSTERING dans [24]). Dans l'ensemble, on peut dire que la complexité de l'algorithme Clust-Agregat proposé est évaluée au pire des cas à  $O(n^2)$ .

## 6. Conclusion

Dans cet article nous avons exploité le fait que différents éléments d'un document XML participent à différentes vues et mènent à des classifications différentes. Cette méthode produit des clusters qui constituent des vues partielles de l'ensemble des données. Nous avons proposé une approche visant à améliorer la qualité des clusters obtenus en exploitant la notion d'agrégation.

Notre approche est basée sur un processus d'optimisation minimisant le désagrément entre des classifications obtenues suite à l'application de l'algorithme k-means. La qualité des clusters obtenus est garantie par le processus d'optimisation d'une part et de la référence de l'algorithme k-means d'une autre part.

Dans le même cadre nous envisageons une étude visant à agréger des algorithmes de classification tels que k-means et POBOC [25]. Nous envisageons une implémentation de cet algorithme sur la base de données Iris pour prouver son efficacité et la différence avantageuse que peut porter notre algorithme par rapport à K-means.

## 7. References

1. Vercoestre et al 2004: Expériences de classification d'une collection de documents XML de structure homogène. 2004 Washington, DC, USA, 2002. IEEE C.S.
2. Mitchell 1999: Machine Learning de Tom Mitchell, New York, McGraw Hill.
3. Guillaume et al 2000: Clustering of XML documents. Computer Physics Communications, 127(2-3):215–227, 2000
4. Denoyer and al. 2003: Structured Multimedia Document Classification, In DocEng'03, Grenoble, France, pp 153-160.
5. Despeyroux et al 2005: Experiments in clustering homogeneous xml documents to validate an existing typology. In Proceedings of the 5th International Conference on Knowledge Management (I-Know), Vienna, Autriche, July 2005.
6. Lee et al 2002: Clustering XML schemas for effective integration. In CIKM '02: Proceedings of the eleventh international conference on Information and knowledge.
7. Steinbach et al 2000: A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.
8. Bertino et al 2001: Measuring the Structural Similarity among XML Documents and DTDs. Technical report, DISI-TR-02-02.
9. Flesca et al 2002: S. Detecting Structural Similarities between XML Documents. In WebDB, pages 55–60, 2002.
10. Nierman et al 2002: Evaluating Structural Similarity in XML Documents. In Proceedings of the Fifth International Workshop on the Web and Databases (WebDB 2002), Madison, Wisconsin, USA, June 2002 .
11. Lian et al 2004: An Efficient and Scalable Algorithm for Clustering XML Documents by Structure. IEEE Trans. Knowl. Data Eng., 16(1):82–96, 2004.
12. Termier et al 2002: A First Step towards XML Data Mining. In ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining, page 450.
13. J. McQueen, 1967: Some methods for classification and analysis of multivariate observations. In the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297.
14. Doucet et al 2002: Naive Clustering of a large XML Document Collection. In INEX Workshop, pages 81–87.
15. Yi et al (2000): A classifier for semi-structured documents, In Proc. Of the 6th International Conference on Knowledge Discovery and Data mining, pp340-344.
16. R. Duda and P. E. Hart 1973: Pattern Classification and Scene Analysis. John Wiley & Sons.
17. K. Fukunaga (1990): Introduction to Statistical Pattern Recognition. Academic Press, San Diego, CA.
18. Strehl and Ghosh 2002: Cluster ensembles: A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research.
19. Fern and Brodely 2003: Random projection for high dimensional data clustering: A cluster ensemble approach. In ICML.
20. Fred and Jain 2002: Data Clustering using evidence accumulation. In ICPR.
21. Filkov and Skeina 2003: Integrating microarray data by consensus clustering. In International Conference on tools with Artificial Intelligence.
22. Topchy et al 2004: A mixture model of clustering ensembles. In SDM.
23. Boulis and Ostendorf 2004 : Combining multiple clustering systems. In PKDD.
24. Arisitde et al 2005: Clustering Aggregation". International Conference on Data Engineering (ICDE), 2005.
25. Cleuziou 2004: Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information Thèse de doctorat, Université d'Orléans. 2004.

## **Une Approche Cognitive Centrée Utilisateurs pour l'Elicitation des Besoins dans les Organisations Distribuées**

Amroune Mohamed<sup>1</sup>, Zarour Nacereddine<sup>2</sup>

<sup>1</sup>Amroune Mohamed, Centre Universitaire de Khenchela-Algérie  
medamroune@gmail.com

<sup>2</sup>Zarour Nacereddine, Laboratoire LIRE, Université Mentouri  
Constantine, 25000, Constantine - Algérie  
nasro-zarour@umc.edu.dz

**Résumé:** L'élicitation des besoins pour un système informatique est une activité conceptuelle complexe dans le processus d'ingénierie des besoins, nécessitant des efforts importants de mise en œuvre, se situant entre le maître d'œuvre et le maître d'ouvrage. Cette phase nécessite une attention particulière surtout quand les acteurs impliqués dans le processus d'élicitation sont géographiquement dispersés dans l'espace et dans le temps. Le travail que nous présentons dans ce papier, s'intègre dans le domaine de l'ingénierie des besoins des systèmes d'information coopératifs.

L'informatisation a conduit au développement, dans les organisations, d'un nombre considérable de systèmes d'information chargés de stocker, d'organiser, d'interroger et de restituer des données utiles à l'utilisateur. Ces systèmes doivent collaborer afin de permettre une coopération à différents niveaux. Les solutions répondant à ces besoins se situent dans le paradigme des systèmes d'information coopératifs.

C'est dans ce contexte que se situe l'approche élaborée, laquelle consiste à développer une approche cognitive centrée utilisateurs pour l'élicitation des besoins dans les organisations distribuées lesquelles sont supportées par les systèmes d'information coopératifs. Cette approche prend en compte les contraintes émergeantes de la dispersion des acteurs dans l'espace et dans le temps.

**Mots Clés:** Ingénierie des Besoins, Elicitation des Besoins, Systèmes d'Information Coopératifs, Groupware, Style Cognitif.

### **1. Introduction**

L'Ingénierie des Besoins (IB) est le processus qui s'intéresse à la première activité dans le cycle de développement de systèmes logiciels. Le processus d'IB inclut généralement les activités de découverte, de spécification, de négociation et de validation de besoins.

Ce travail se situe dans le contexte où une organisation souhaite atteindre un nouvel objectif, mais n'ayant pas toutes les compétences nécessaires et/ou ne disposant pas de ressources (humaines, technologiques et financières) suffisantes pour réaliser seule cet objectif, elle décide de coopérer avec d'autres partenaires. Ainsi, cette coopération génère la construction d'organisations distribuées lesquelles sont supportées par des Systèmes d'Information Coopératifs (SIC) [Zarour, 2004]. Une fois les partenaires sélectionnés, ils doivent demander à une équipe spécialisée dans l'IB d'établir un cahier de charges du travail Coopératif.

Deux espaces de recherche essayent de trouver des solutions à ces problèmes. Le premier espace utilise le domaine de travail coopératif assisté par ordinateur (littéralement CSCW) [Domain and Zoughi, 2002]. Le second espace utilise l'informatique cognitive [Chiew and Wang, 2003]. Notre contribution s'intègre dans le deuxième espace de recherche (informatique cognitive).

Concrètement, cet article décrit une approche cognitive, centrée utilisateurs et axée beaucoup plus sur la phase d'élicitation des besoins dans les organisations distribuées où chacune est supportée par un système d'information.

### **2. Ingénierie des besoins**

Cette partie décrit un résumé des principales approches dans le domaine de l'IB : les approches dirigées par les objectifs (buts), les approches à base de scénarios et les approches couplant les deux premières approches citées.

Les approches d'IB orientées objectifs (buts) considèrent les buts comme un potentiel motivant à l'action et se diffèrent des plans, des procédures ou d'autres moyens d'atteindre le but. Nous citons comme exemple l'approche I\* [Castro and Kolp, 2002].

Les approches à base de scénarios font l'hypothèse qu'il est plus facile de décrire, le plus souvent dans un récit en langue naturelle, ce qui se passera dans le système futur que d'exprimer directement les objectifs du futur système. L'émergence des scénarios d'usages (use cases) [Jacobson et al., 1993] dans le monde de l'orienté objet confirma cette tendance. Mais, Les approches à base de scénarios présentent elles aussi quelques limites [Lamsweerde, 2000].

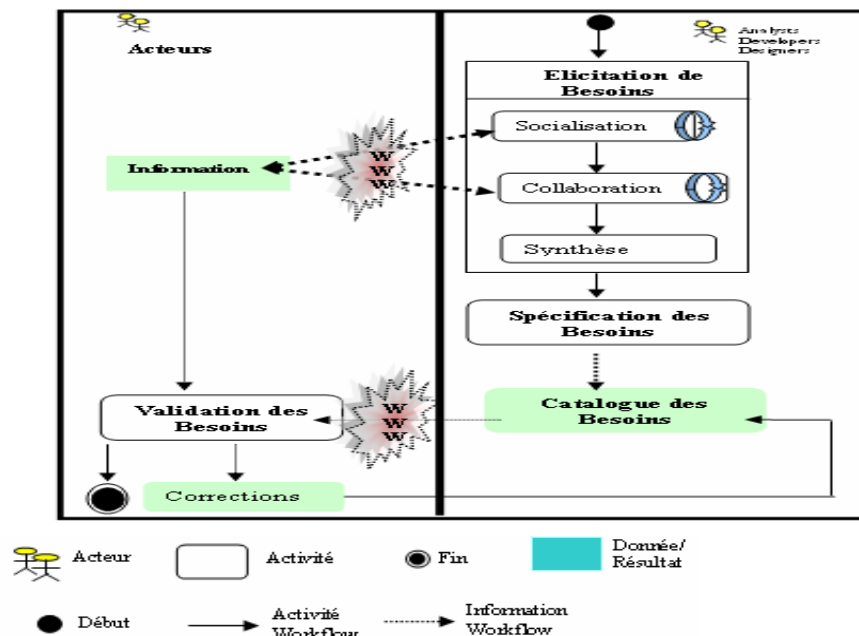
Afin de pallier aux difficultés recensées dans les approches orientées buts et les approches à base de scénarios, les approches couplant buts et scénarios.

En effet, les trois approches présentées dans cette section ne peuvent pas être utilisées telles quelles dans le contexte des organisations distribuées puisqu'elles ne prennent pas en considération tous les nouveaux paramètres (décalage horaire, etc.) intrinsèques à ces organisations. D'où l'objectif de ce travail où nous présentons dans la section suivante une approche qui prend en charge ces paramètres dans le contexte des SIC.

### 3. L'Approche Proposée

Les problèmes connus qui s'opposent au processus d'IB distribués, sont [Domain and Zoughi; 2002], [Betty H.C. Cheng, et Joanne M. Atlee ;2007] : (1) Différences des horaires (décalage horaire). (2) Différences de culture (surtout quand les acteurs sont de différentes nationalités). (3) Diversité culturelle professionnelle (ou de métier) dans le cadre d'une coopération entre les organisations de métiers différents où elles possèdent des cultures d'organisation différentes (procédés, modes de management, moyens et technologies, modèles organisationnels, etc.). (4) Le choix des outils adéquats de communication à distance. Et (5) Les problèmes de motivation, d'engagement et de confiance entre acteurs.

L'approche que nous proposons est itérative (figure 1). Elle est composée de trois étapes essentielles : (i) *La phase d'élicitation (ou découverte) des besoins*, (ii) *La phase de spécification*, (iii) *La phase de validation*.



**Figure 1 : Les phases d'IB de l'approche proposée**

Les étapes proposées sont inspirées des travaux de [JOSÉ E, NORA K, 2004]. Notre contribution par le biais de l'approche proposée est d'adapter les phases de l'approche décrite dans (JOSÉ E, NORA K) dans un contexte d'IB coopératifs distribués. L'adaptation du processus concerne essentiellement la phase d'élicitation des besoins qu'on considère comme étant la plus importante des phases, du fait que le résultat

de cette phase est utilisé par toutes les autres phases. L'approche proposée décompose la phase Elicitation en trois sous phases (figure 1): (1) *phase de socialisation*, (2) *phase de collaboration*, (3) *phase de synthèse*.

Nous désignons par (WWW) dans la figure 1 que la communication entre les acteurs et les différents membres du groupe de l'IB est assurée par les moyens d'un «groupware».

### 3.1 Phase de Socialisation

Les principaux objectifs de cette phase sont la définition des acteurs, l'utilisation des styles cognitifs pour bien connaître les profils des acteurs et enfin la prise de connaissance entre ces acteurs qui sont membres de l'organisation virtuelle.

Cette phase commence par la définition des acteurs impliqués dans le projet de la coopération.

La communication entre acteurs implique des aspects des mécanismes humains. Ces aspects sont analysés dans la science cognitive. Dans ce cas, nous sommes intéressés beaucoup plus par quelques techniques de psychologie appelées les "Modèles de Style d'Apprentissage" (Learning Style Models) [Hickey and Davis, 2003].

Ces modèles sont utilisés pour bien comprendre les rapports Enseignants/Etudiants dans un système de e-learning, et comme dans la phase élicitation des besoins, chaque acteur (y compris l'analyste) apprend des autres acteurs, donc un acteur dans une situation est vu comme « étudiant » et dans une autre, il est considéré comme « enseignant ». En partant de ce point de vue, nous allons tirer profit en adaptant ces modèles cognitifs dans la phase d'élicitation de notre approche, afin de comprendre le rapport Analyste/Acteurs.

Pour bénéficier des modèles des styles cognitifs dans notre approche, nous utilisons le modèle de style cognitif proposé par Felder-Silverman [Felder and Silverman, 1988]. Ce modèle classe les personnes (dans notre cas les acteurs) en quatre grandes catégories où chaque catégorie est subdivisée elle aussi en deux sous catégories : *Actif/Réflexif*, *Analysateur/Intuitif*, *Visuel/Verbal* et *Séquentiel/Global*.

Pour classer un acteur dans une catégorie, un test de choix multiple (questionnaire<sup>1</sup>) est considéré. La préférence envers une catégorie ou une autre peut être forte, modérée ou faible. Seulement quand il y a une préférence forte, une personne est classée en tant que membre de ce groupe. L'analyste peut éventuellement choisir les outils du groupware adéquats aux caractéristiques des acteurs selon leurs styles cognitifs. A titre d'exemple nous pouvons dire que les moyens de la visioconférence s'adaptent bien pour les acteurs visuels.

Le deuxième objectif assigné à cette première sous phase est la prise de connaissance entre les différents acteurs. Des recherches ont mis en évidence que l'identification à un groupe améliore l'engagement individuel dans la réalisation d'un objectif commun [Paulus & al, 2002].

Dans ce même cadre d'idée, l'approche proposée prévoit une séance de travail collectif afin de choisir un nom pour l'organisation virtuelle créée, à ce niveau, et en cas de conflit, le recours à une opération de vote est utilisé comme moyen d'arbitrage.

Après avoir terminé la phase de socialisation, débutera la deuxième phase qui est celle de Collaboration. La section suivante décrit en détail cette phase.

### 3.2 Phase de Collaboration

Cette phase consiste à éliciter les besoins des acteurs. Elle se décompose elle-même en plusieurs séquences: (1) *Brainstorming Electronique*, (2) *Classement individuel des Besoins* et (3) *Débat et classement collectif des besoins (Catégorisation des Besoins)*.

#### 3.2.1 Brainstorming Electronique

Le groupe ainsi formé de l'analyste et des différents acteurs des différentes organisations, doit donner dans un forum de discussion toutes les idées, pensées et intuitions qui leur viennent à l'esprit pour définir leurs besoins (exigences).

Dans une séance de brainstorming la participation des acteurs peut être manifestée en utilisant ce qu'on appelle « fonctions coopératives » marquant le type d'interaction entre acteurs, Comme exemple de ces fonctions nous citons : Poser une question, Répondre à une question, Exprimer son point de vue,

<sup>1</sup> <http://www.engr.ncsu.edu/learningstyles/ilsweb.html>

Proposer une sous discussion, Demander un complément d'information, Chercher à convaincre, Défendre son point de vue. Les séances de brainstorming sont suivies par un classement individuel des différentes idées recensées (ce travail est effectué au niveau de chaque organisation hors ligne).

### 3.2.2 Classement Individuel des Besoins

Les acteurs de chaque organisation doivent organiser individuellement et hors ligne les différentes exigences et de les classer de la plus à la moins centrale pour réaliser une tâche collective. Cette séquence de travail individuel est jugée intéressante, en se basant sur les résultats de recherche [Paulus & al, 2002] qui soulignent l'intérêt de faire suivre une séance collective (séance de brainstorming électronique) d'une séance de travail individuel. En plus cette séance permet de limiter l'effet du faux consensus généralement observé lors des séances du travail collectif.

### 3.2.3 Débat et Classement Collectif des Besoins (Catégorisation des Besoins)

La sous phase de classement individuel est suivie par des séances réservées à un classement collectif des différents besoins dégagés. Cette opération consiste à classer les différents besoins en un ensemble de catégories de besoins; les trois grandes catégories sont (*Très Importantes, Importantes, Accessoires*).

## 3.3. Phase de Synthèse

Après avoir terminé toutes les séances asynchrones et synchrones, l'analyste doit procéder à synthétiser tout le travail, il peut utiliser pour cela toutes les données recueillies des différentes séances de travail en ligne. De multiples sources de données peuvent être utilisées: (a) Les observations du groupe (Synchrone et Asynchrone), (b) Questionnaires et (3) Entretiens avec les participants.

Dans le cadre de la synthèse, l'analyste peut servir de la méthode CREWS (Cooperative Requirements Engineering With Scenario) pour éliciter des nouveaux buts, que le processus précédant n'a pas pu les éliciter.

Le résultat de la phase Elicitation est un document présentant toutes les exigences exprimées lors de cette phase.

L'analyste doit classer les besoins fonctionnels et non fonctionnels et essayer d'établir un ordre de priorité entre les besoins fonctionnels en se basant sur la méthode AHP [Ahl, 2005].

## 4. Etude de Cas

Notre approche est appliquée sur un exemple de *développement d'un Progiciel de Gestion Intégrée (PGI), en anglais ERP<sup>2</sup>*, faisant appel à la coopération de plusieurs organisations spécialisées chacune dans un domaine bien spécifié tels que la gestion comptable et financière, la gestion de la production et de la logistique, la gestion des ressources humaines, la gestion administrative ainsi que la gestion des ventes et des achats. Nous supposons que ces organisations sont géographiquement dispersées. Donc la première tâche importante est l'élaboration d'un cahier de charges pour le développement de l'ERP.

Pour le déroulement des phases de l'approche, nous envisageons dans cet article uniquement trois aspects, chacun relevant d'une phase. Dans le cadre de la définition des acteurs, l'extrait suivant du questionnaire, relatif aux styles cognitifs est donné : Chaque acteur est censé de répondre à ce questionnaire.

<b>Question 01 :</b> ( <i>When I think about what I did yesterday, I am most likely to get</i> )	
	(a) pictures.
	(b) words.
<b>Question 02 :</b> <i>I prefer to get new information in</i>	
	(a) pictures, diagrams, graphs, or maps.
	(b) written directions or verbal information.
<b>Question 03 :</b> <i>remember best</i>	
	(a) what I see.
	(b) what I hear.



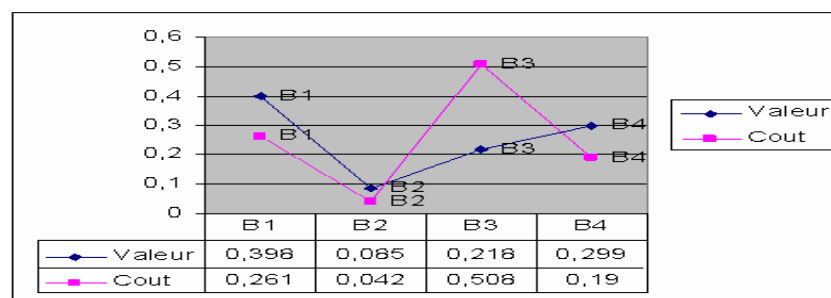
Notons à ce niveau que si la réponse est (a) pour les trois questions, ceci donnera un indice que l'acteur peut appartenir à la catégorie Visuelle.

Dans la phase Collaboration, l'exemple suivant montre un Classement Collectif d'un ensemble de Besoins (relatifs au module finance) selon les trois catégories de besoins (Très importants, importants, Accessoires).

**Exemple :**

Besoins Très Importants	
☞	Permettre à l'utilisateur d'exécuter des requêtes en fonction de critères spécifiques.
☞	Donner la possibilité aux utilisateurs de modifier la hiérarchie de déploiement et la structure d'un plan de comptes.
Besoins Importants	
☞	Assurer la représentation graphique des données financières par le biais d'un outil de génération de rapports internes, ou par l'intégration directe d'outils de génération de rapports tiers.
☞	Le module financier doit assurer la création, le routage et le stockage électroniques des documents.
Besoins Accessoires	
☞	La possibilité d'utiliser Microsoft Excel ou d'autres outils de productivité du type tableur.

Pour la priorisation des besoins, le diagramme Coût-Valeur de la méthode AHP est utilisé pour identifier les besoins de plus grandes valeurs et de moindres coûts de quatre besoins (B1....B4). On peut le considérer comme le besoin le plus prioritaire en tenant compte des deux critères (Valeur-Coût).



**Figure 4.** Diagramme Coût/Valeur

## 5. Conclusion

L'une des causes principales des problèmes rencontrés dans la « crise du logiciel » est due principalement aux manquements dans le processus d'IB. Cette étape doit être menée de manière concise et efficace.

Le développement d'applications permettant à des organisations géographiquement éloignées de travailler ensemble est une tâche difficile. L'approche que nous avons proposée, est structurée en trois phases (Socialisation, Collaboration et Synthèse). Afin d'atteindre les objectifs assignés pour chaque phase, l'approche proposée profite de plusieurs disciplines (IB centralisés, les SIC, le TCAO, le groupware, la conception centrée utilisateur, les méthodes de prise de décision, la science cognitive, etc.). Pour valider l'approche proposée, nous l'avons appliqué sur une étude de cas relative au développement d'un progiciel de gestion intégrée (ERP).

Nous envisageons comme perspectives de notre travail, le développement des phases Spécification et Validation, proposées dans le processus général d'IB dans les organisations distribuées.

## 6. Références

- [Ahl, 2005] Ahl V., An Experimental Comparison of Five Prioritization Methods. Master's Thesis, School of Engineering, Blekinge Institute of Technology, Ronneby, Sweden, 2005.
- [Betty H.C. Cheng, Joanne M. Atlee, 2007] : Research Directions in Requirements Engineering. Future of Software Engineering(FOSE'07) , IEEE 2007.
- [Castro and Kolp, 2002] Castro J., Kolp M., Towards requirements-driven information systems engineering: the Tropos project." Information Systems, Elsevier Science Ltd.27, 2002, pp. 365-389.
- [CREWS, 1999]. CREWS (ESPRIT N°21.903), Cooperative Requirements Engineering With Scenarios, <http://SunSITE.Informatik.RWTH-Aachen.DE/CREWS>.
- [Chiew and Wang, 2003] Chiew, V., and Wang, Y., From Cognitive Psychology to Cognitive Informatics, *Second IEEE International Conference on Cognitive Informatics*, ICCI'03. London, UK, August 2003, pp. 114-120.
- [Domain and Zoughi, 2002] Domain D.E. and Zoughi D., An insight into the interplay between culture, conflict and distance in globally distributed requirements negotiations, 36th IEEE Hawaii International Conference on System Sciences (HICSS'03), 0-7695-1874-5/03, 2002.
- [Felder and Silverman, 1988] Felder, R., and Silverman, L., Learning and Teaching Styles in Engineering Education, *Engineering Education*, 78(7), 1988, pp. 674-681.
- [Hickey and Davis, 2003]. Hickey, A.M. and Davis, A., Requirements Elicitation and Elicitation Technique Selection: A Model for Two Knowledge-Intensive Software Development Processes, 36th Annual Hawaii International Conference on Systems Sciences (HICSS), January 2003, pp. 96-105.
- [José and Nora, 2004] José E. and Nora K., Requirements Engineering for Web Applications – A Comparative Study, *Journal of Web Engineering*, Vol. 2, No.3, Rinton Press, 2004, pp. 193-212.
- [Jacobson et al., 1993] Jacobson I., Christerson M., Jonsson P., and Övergaard G., Object-Oriented Software Engineering, A Use Case Driven Approach. Addison-Wesley, ACM Press, 1993.
- [Lamsweerde, 2000] Lamsweerde A. V., Requirements engineering in the year 00: A research perspective. In *Proceedings 22nd International Conference on Software Engineering*, Invited Paper, ACM Press, June 2000.
- [Mc Graw and Haribson, 1997] Mc Graw K. and Haribson K., User centred requirements, the scenario-based engineering process. Lawrence Erlbaum Associates Publishers, 1997.
- [Paulus & al, 2002] Paulus P.B, Legett-Dugosh K., Dzindolet M.T., Coskun H., & Putman V.L. Social and cognitive influences in group brainstorming: predicting production gains and losses. In W. Stroebe, & M. Hewstone (Eds.), *European Review of Social psychology*, London Wiley, 2002, pp. 299-325.
- [Yue, 1987] Yue K., What does it mean to say that a specification is complete? *Proc. IWSSD-4. Four International workshop on software specification and design*, Monterrey, 1987.
- [Zarour, 2004] Zarour N., Contribution à la modélisation de la coopération des systèmes d'information distribués et hétérogènes: Le système DAARACHE- Application aux Entreprises de production. Thèse présentée pour l'obtention de grade de docteur d'Etat en Informatique, université Mentouri de Constantine, Septembre 2004.

## Relevé et classification automatique des dégradations des routes bitumées

Guellouma Younes<sup>1</sup>, Hadda Cherroun<sup>1</sup>,

<sup>1</sup> Université Amar Telidji Laghouat  
37G Route de Ghardaia Laghouat 03000, Algérie  
{y.guellouma, hadda\_cherroun}@mail.lagh-univ.dz

**Résumé:** Le but de cet article est la conception d'une technique efficace pour le relevé automatique des dégradations sur une route bitumée. L'objectif principal de notre étude est de pouvoir identifier des dégradations sur un flux vidéo. Pour ce faire, nous nous sommes intéressés dans un premier temps au problème de classification d'images de dégradations, et ceci dans l'espoir de pouvoir généraliser le procédé au flux vidéo. Le problème de classification d'images a été très étudié, notamment dans le domaine de l'imagerie médicale. Par rapport à cette dernière, notre problématique rajoute principalement deux difficultés. Le premier est la caractérisation des différents types de dégradations (nid de poule, affaissement, ...) et la deuxième est la suppression du bruit (trace de pneu, tâche d'huile, ...). Dans cet article, nous donnons une caractérisation d'un certain type de dégradations, puis nous proposons une méthode basée sur les techniques d'apprentissage pour la classification des images de dégradation. Nous terminons par des questions ouvertes sur le problème du relevé automatique de dégradation ainsi que des perspectives.

**Mots clés:** Dégradation, Plus proche voisin, Traitement d'images, Segmentation, Classification, Data mining.

### 1. Introduction

La maintenance du réseau routier d'un pays est parmi les tâches les plus importantes pour la prospérité de son économie. Vu la taille gigantesque de ce support principal de transport, le suivi automatique ou semi-automatique de la maintenance devient inévitable.

À partir d'images numériques, la déduction ou la reconnaissance d'une dégradation depuis une image traitée est une tâche très difficile car une image de la chaussée peut porter plus d'informations que celles des dégradations qui peuvent tromper le processus de reconnaissance.

Dans cet article, nous proposons une approche pour la résolution de ce problème. Elle est basée sur des techniques de traitement d'images ainsi que sur les méthodes de classification automatique.

Dans la première section, nous présentons les différentes notions utiles à la compréhension du problème de Relevé Automatique des Dégradations (RAD).

Ensuite, nous présentons les différentes étapes de notre approche pour la résolution du problème RAD. Enfin, nous terminons par quelques problèmes issus du RAD ainsi que des perspectives pour l'amélioration de notre approche.

### 2. Le relevé automatique des dégradations

Dans cette section, nous commençons par définir quelques notions utiles à notre étude. Ensuite, nous rappelons quelques techniques classiques de relevé des dégradations. Enfin, nous présentons les différentes difficultés issues du RAD à partir d'images numériques.

Les dégradations de routes peuvent être de différents types, les experts des travaux publics recensent plus de 16 types. Les plus importants :

- Les nids de poules : Un nid-de-poule est une cavité dans la chaussée aux abords découpés qui se crée lorsque le revêtement routier s'effrite et que les matériaux de remblais se dispersent. Les nids de poule sont généralement causés par le gel et le dégel successif des routes, à une mauvaise qualité d'enrobé mis en place ou à des problèmes de structures de chaussées.

- Les affaissements : Le phénomène d'affaissement se produit lorsque les terrains sont plutôt plastiques et que la profondeur d'exploitation est importante par rapport à l'épaisseur de la taille. Les affaissements sont les dégradations les plus difficiles à détecter car ils ne sont pas clairs sur une image 2D, surtout si l'affaissement occupe une surface importante.

## **2.1 Relevé des dégradations**

Le relevé des dégradations d'un réseau routier constitue la phase la plus importante dans son entretien, en effet, le relevé des dégradations compote la collection et la constitution d'un inventaire du terrain à l'aide d'un ensemble de moyens et le traitement et l'importation des données collectées dans les outils du gestionnaire [1, 2].

Les techniques traditionnelles du traitement des dégradations se font d'une manière manuelle; une équipe d'experts effectue des déplacements périodiques, fait le diagnostic du type de dégradation et propose des solutions de maintenance en conséquence.<sup>3</sup>

Récemment, d'autres techniques de relevé sont utilisées ; dans certains pays, il existe un autre moyen utilisé : un grand engin équipé d'un émetteur/récepteur de signaux sonores infrarouges ou laser, est capable de détecter les dégradations même dans les couches inférieures de la chaussée lors de son parcours [3].

Cependant une tendance vers le traitement automatique des dégradations a vu le jour depuis quelques années se basant essentiellement sur les techniques de traitement d'image. Par exemple, pour le compte du ministère du transport français, le groupe d'informatique et des technologies de l'information a mis au point un logiciel (PDDC) qui consiste en une application sur système embarqués utilisée par les préposés officiels lors des inspections sur le terrain. Le programme PDDC est conçu de façon à intégrer les données complètes se rapportant à différents états de dégradation de la chaussée; parmi celles-ci, mentionnons les mesures du degré de sévérité et de densité, les fiches d'entretien antérieures, ainsi que les calculs portant sur l'état de dégradation général [4]. Nous pouvons classer ce système comme système d'aide à la décision.

## **2.2 Problèmes induits par le RAD**

Le relevé et la classification des images des dégradations à partir d'images numériques voir même le flux vidéo, peuvent être réalisés d'une manière automatique en faisant appel aux techniques de traitement d'images et de classification automatique.

Le problème de classification automatique d'images a été très étudié, notamment dans le domaine d'imagerie médicale [5]. Cependant, le problème RAD rajoute d'autres difficultés particulières à savoir :

- La détection des fausses dégradations :
  1. La couleur du goudron : l'utilisation de plusieurs matières dans la couche extérieure de la chaussée donne une couleur non homogène (plusieurs couleurs apparaissent sur une chaussée saine).
  2. Les dégradations déjà traitées : une dégradation traitée ne prend plus la couleur dominante, cela mène à une image avec plusieurs surfaces différemment colorés.
  3. Les traces des pneus, d'huile : les pneus et l'huile et d'autres sources peuvent laisser des traces qui changent de couleur par rapport au reste de la chaussée, il ne faut pas considérer ses tâches comme étant des dégradations.
- L'apparence des dégradations : une dégradation n'a pas une seule forme et une seule couleur mais elle peut en prendre plusieurs, cela complique la reconnaissance de ces formes (par exemple dans une image satellitaire, il est facile de reconnaître les formes : l'agglomération a toujours une couleur très claire ou blanchâtre, les rivières sont bleuâtres et sont définis en largeur et en longueur..., par contre cela ne paraît pas évident dans une dégradation de chaussée).
- Déduction des informations de la troisième dimension : une dégradation de la chaussée provoque aussi bien des déformations en profondeur qu'en largeur, mais il est difficile de déduire ces informations sur une image 2D.
- Le reflet de l'environnement : pour une seule dégradation, une image prise au soleil, et une autre en pleuvant, une le matin et une l'après midi ne donne absolument pas les mêmes couleurs (Contraste, reflets et ombres).

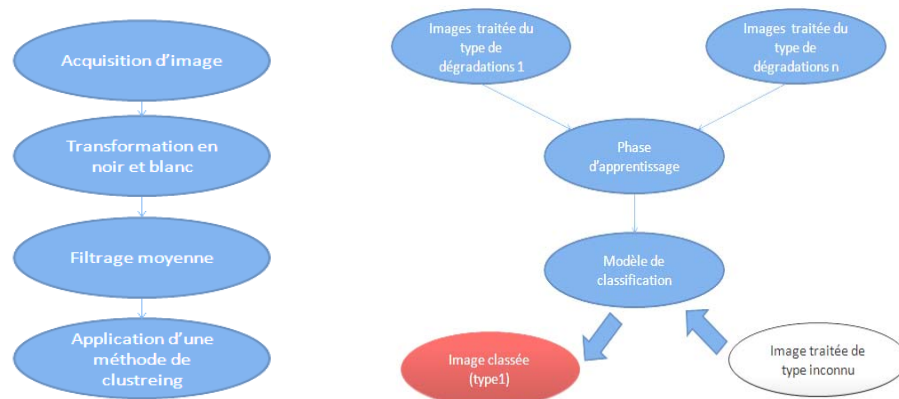
---

<sup>3</sup> Selon l'expert de la direction des travaux publics de la wilaya de Laghouat interviewé le 27 /10/2007

Alors pour résoudre le problème RAD à partir d'images numériques, il faut trouver une approche qui est capable de classer chaque image dans le bon type de dégradation tout en gérant les difficultés citées, dans la prochaine section, nous essayerons de définir une approche basée sur des méthodes de traitement d'images ainsi que des techniques de représentation et de classification automatiques de données.

### 3. Nouvelle approche

L'approche proposée comporte deux étapes (Fig 1), la première est une chaîne d'opérations de traitement d'image qui a pour rôle de préparer l'image pour en extraire des informations utiles. La deuxième étape consiste en l'application d'une méthode de classification d'image (d'apprentissage supervisé) en se basant sur l'un des algorithmes de classification en faisant appel aux techniques les plus utilisées dans le domaine du data mining.



**Fig 1.** Les différentes étapes de l'approche proposée pour la résolution du RAD

- Le prétraitement d'image : L'acquisition d'images Il est préférable que les images soient acquises par l'intermédiaire d'un appareil numérique, dans le but d'éviter le bruit. L'échelle de l'image doit être aussi enregistrée, quoiqu'il est meilleur d'utiliser des images à échelle unie.
- Transformation en niveaux de gris Une image numérique en couleur est codée généralement en RVB (pour Rouge, Vert et Bleu) qui utilise 24 bits dont chaque couleur est représentée sur un octet, soit 224 couleurs (plus de 16 millions de couleurs), or qu'une image en niveaux de gris comporte seulement 256 niveaux de gris (un seul octet) [6], alors il est plus aisé d'utiliser des images en niveaux de gris que des images couleurs, cette phase est très utile dans le traitement des dégradations car une image du genre contient toujours des couleurs très proches, la transformation en niveaux de gris portera toujours l'information.
- Le filtre moyenne: Il y a plusieurs méthodes de filtrage utilisées dans le domaine de traitement d'images ; Le filtrage consiste à appliquer une transformation (appelée filtre) à tout ou partie d'une image numérique en appliquant un opérateur [7]. La surface du goudron est –malgré qu'elle a une couleur dominante- est une surface qui contient beaucoup de grains (à cause de la matière de construction), pour réduire le taux des grains et offrir plus d'homogénéité aux surfaces, le filtre moyenne dont le principe est de faire la moyenne des valeurs des pixels avoisinants peut donner un bon résultat.
- Application d'une méthode de clustering: La phase la plus importante dans le prétraitement d'images est la segmentation d'image [8], elle donne une image avec un nombre réduit de surfaces (segments) dont chacune représente une partie de l'image possédant des caractéristiques communes. Il y a plusieurs méthodes utiles pour la segmentation d'image tel que le seuillage multiple [9], la segmentation par champ de Markov [10], la segmentation par champ de Gaib [10]. Les méthodes de clustering largement utilisées dans le domaine du data mining ont prouvé leurs efficacité dans la segmentation d'image car ces méthodes n'ont pratiquement pas besoin d'informations sur les images, de plus elles peuvent détecter les régions similaires selon les techniques utilisées dans chacune.

le « K means » (appelé aussi C moyenne ou méthode des centres mobiles) est un algorithme de clustering non supervisé largement utilisé dans le data mining .Le K means est devenu un algorithme très utilisé dans le traitement d'images à cause de sa capacité de détecter les groupements de surfaces

similaires sans l'obligation de définir des contraintes sur les couleurs ou les formes suspecte d'apparaître, cet algorithme a fait ses preuves dans plusieurs domaines à noter la détection des tumeurs cérébrales, la reconnaissance de texte, etc. Le K means se base sur la notion de distance entre les enregistrements, pour cela il utilise plusieurs distances connues (Euclidienne, Manhathan ....) [11, 12]. Pour les images, en niveaux de gris, la distance sera largement plus facile à écrire :

$$D(Pixel_1, Pixel_2) = G_{P_2} - G_{P_1}$$

Où  $G_{PX}$  est le niveau de gris du  $Pixel_X$ . Pour segmenter l'image en K surfaces, les étapes de l'algorithme, décrit en détails dans [12], sont :

1. A partir d'un échantillon de P enregistrements (image de  $m \times n = P$  pixels).
2. Choisir K centres initiaux  $c_1 \dots c_K$ .
3. Repartir chaque enregistrement de P dans un cluster i dont le centre  $c_i$  est le plus proche.
4. Si aucun enregistrement ne change de cluster l'algorithme est terminé.
5. Sinon calculer les nouveaux centres  $c_i$  et aller à 3.

Le calcul de nouveaux centres  $c_i$  se fait selon la variante utilisée de l'algorithme K means. Dans le K means classique, le nouveau centre  $c_i$  est la moyenne des éléments du cluster  $i$ . (la moyenne des niveaux de gris).

Nous rajoutons une étape supplémentaire, à la fin de l'algorithme de K means, chaque pixel  $i$  appartenant à un cluster  $j$  prendra la couleur du centre  $C_j$ . La Fig. 1, schématise ce que l'on obtient à partir de notre image du nid de poule cité précédemment. En effet, cette étape donne à chaque cluster la couleur dominante dans la surface représentée, cela ajoute une information supplémentaire pour les étapes suivantes.

### 3.1 La classification automatique

Il est très difficile de définir des contraintes sur lesquels se base la classification des dégradations, (vu la variété de formes que peut prendre une seule dégradation). De plus on ne peut pas en être sûr que les caractéristiques définies pour chaque dégradation sont suffisantes pour décider qu'une telle image présente vraiment une telle dégradation (on sait par exemple qu'un nid de poules a toujours une forme connexe et régulière, mais l'ordinateur peut se tromper si des fissures constituent une forme connexe et régulière).

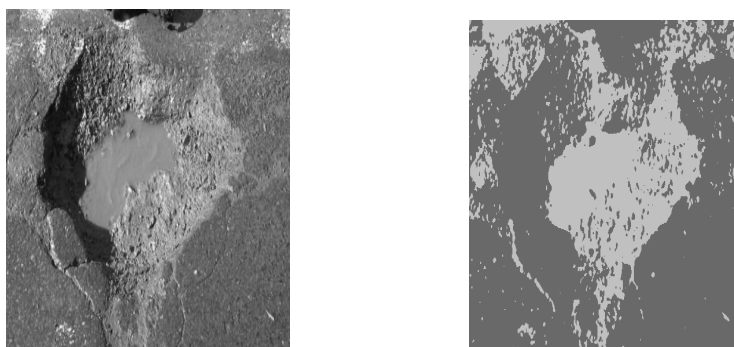


Fig. 1 : Image segmentée par le k means

L'idée de l'approche définie est qu'à partir d'exemples déjà résolus qui représentent la connaissance (pour chaque type de dégradation, on apporte un échantillon d'images variées), la décision de classer une nouvelle image est prise ; pour cela, plusieurs étapes sont menées :

- La représentation d'image : Les méthodes de classification automatique utilisent des échantillons avec des valeurs quantitatives (numériques dans la plus part des cas) [14], il faut trouver une bonne représentation pour les images afin de passer aux méthodes de classification, nous proposons d'utiliser l'analyse en composantes principales ACP, l'ACP fournit une méthode de représentation d'une population décrite par un ensemble de caractères dont les modalités sont quantitatives. L'ACP permet de structurer une population dans la mesure où elle permet de "visualiser" des regroupements

d'individus, de construire des variables synthétiques, "résumant" l'ensemble des variables prises en compte dans l'analyse.

L'utilisation de l'ACP sur la représentation proposée remplace la matrice de dimension  $(n, m)$  par  $K$  vecteurs à une dimension, (on peut remplacer la matrice par un seul vecteur). Ces vecteurs qui « compressent » la matrice gardent toujours l'information portée par la représentation de l'image [15, 16].

- La phase d'apprentissage : La phase d'apprentissage consiste à « apprendre » au système comment reconnaître une dégradation depuis une image, pour cela, un ensemble de connaissances doit être présent. Cet ensemble est constitué de caractéristiques définies pour chaque type de dégradations plus un ensemble d'images déjà classées.

Après les limites des systèmes experts, d'autres méthodes ont fait un grand avancement, dont les plus utilisés sont les méthodes de classification (appelés aussi méthodes d'apprentissage supervisé) largement utilisés dans le domaine du data mining, dans notre approche, on va utiliser l'algorithme du plus proche voisin [17]. Parmi ces méthodes, la méthode du plus proche voisin (KNN pour  $K$  nearest neighbours) est une méthode très utilisée dans le classement automatique du data mining, elle est basée sur la notion de ressemblance, chaque nouvel individu est comparé avec d'autres individus déjà classés, la décision se fait selon la ressemblance du nouvel individu par rapport aux autres [18,19].

La phase d'apprentissage dans la méthode KNN n'est rien qu'une constitution d'un espace d'exemples résolus (images classés par l'expert) qui forment ce qu'on appelle « modèle d'apprentissage ». A partir de ce modèle, les nouvelles images vont se classer tout en reposant sur la notion de similarité.

Pour cela l'ensemble d'images d'apprentissage, ayant subi le prétraitement défini en haut, doivent être représentées d'une manière numérique afin que l'algorithme KNN puisse être appliqué. Rappelons le principe de l'algorithme [19]:

1. L'image via sa représentation est comparée aux autres images d'apprentissage (déjà classées). Les  $K$  plus proches images seront acceptés comme voisinage.
2. Chaque image du voisinage fait un « vote » pour la classe associée.
3. L'image traitée appartient à la classe majoritaire (qui a plus de vote).
4. Un degré de « confiance » est affecté à la décision, la confiance est égale à la différence entre nombre de votes pour la classe majoritaire et le nombre total des votes.

### 3.2 Complexité de l'approche

L'approche proposée est constituée de plusieurs étapes, chaque étape nécessite un temps d'exécution et donne un résultat à la prochaine étape, malheureusement, il est difficile de donner une complexité exacte pour notre approche à cause de la multitude de paramètres qui influent sur le processus de classification, nous essayerons de donner une idée sur cette complexité. Si on prend  $X$  la taille de l'échantillon d'apprentissage et  $(m, n)$  les dimensions moyennes de l'image, chaque  $X_i$  doit subir la chaîne de prétraitement dont :

La transformation en niveaux de gris prend  $O(m \times n)$  (avec  $m$  et  $n$  sont les dimensions d'une image) comme complexité.

Le filtre s'exécute en  $O(m \times n)$ .

Le  $K$  means est un algorithme relativement lent, car il recalcule à chaque itération, les distances entre chaque pixel et chaque centre de clusters choisis, mais l'avantage de cet algorithme réside de sa convergence assez rapide. Alors si on prend  $P = m \times n$  la taille de l'image (dont  $n$  est le nombre de pixels en hauteur et  $m$  le nombre de pixels en largeur),  $K$  le nombre de clusters et  $T$  le nombre d'itérations, la complexité théorique est d'ordre  $O(P \times K \times T)$ .

L'algorithme Knn est très gourmand en temps car à chaque fois qu'une image doit être classée, l'image doit être comparée avec toutes les images d'apprentissage, des calculs de l'ordre matriciel (différence entre matrices, ordonnancement de matrices...), on estime sa complexité temporelle en  $O(m \times n \times p \times k)$  avec  $n$  et  $m$  les dimensions de matrice représentante,  $p$  le nombre d'images d'apprentissage et  $k$  le nombre de matrices (images sélectionnées) pour la décision.

#### **4. Conclusion et Perspectives**

Nous avons conçu une nouvelle approche pour la résolution du problème RAD. Notre méthode se base sur des techniques de traitement d'images ainsi que sur des méthodes de classification supervisée efficaces, chaque image subi une chaîne d'opérations de traitement d'images allant de la transformation de l'image en niveaux de gris, l'application d'un filtre passe bas (moyenne) jusqu'à segmenter l'image en utilisant la méthode du K means, puis la classification se fait en utilisant la méthode du plus proche voisin (Knn).

Cette approche représente un noyau pour la résolution du problème de RAD qui doit être validée par une expérimentation sur des données réelles du domaine. Une généralisation de cette méthode à d'autres domaines nécessitant l'appel à des traitements d'image spécifiques.

#### **5. Références :**

- [1] Info DLC Quebec: Spécifications d'uni et traitement des données d'IRI, Vol. 4, n° 4, Avril 1999.
- [2] Geo Invent Sud: Relevé routier 2005, lien: [http://geoinvent.fr/releve\\_routier.htm](http://geoinvent.fr/releve_routier.htm)
- [3] Geo Invent Sud: Prestation d'auscultation des chaussées 2005.
- [4] Road Talk Volume 11 n°3, ministère de transports français, automne 2005 : Le système PDDC et les inspections automatisées.
- [5] Z.F. Knops, J.B.A. Maintz, M.A. Viergeve, J.P.W. Pluim. Medical Normalized mutual information based registration using *k*-means clustering and shading correction. Mars 2005.
- [6] Nicolas janey: Le traitement d'images. Cours de Master en infographie 2008.
- [7] Isabel Bloch : Filtrage d'images. Département TSI signal-image. Novembre 2007
- [8] Max Mignotte : Segmentation d'images sonar par approche markovienne hiérarchique non supervisée et classification d'ombres portées par modèles statistiques. 1998.
- [9] G Delyon, P Martin, F Guerolt. Segmentation d'images par minimisation de complexité stochastique. Octobre 2004.
- [10] Jian Feng Yao : Analyse d'images par champ de Markov. 2006.
- [11] Omar El Ganaoui & Mathieu Perrot : Segmentation par la méthode des K means. Décembre 2004.
- [12] A Moussaoui : Clustering data mining. Cours école doctorale ingénierie des systèmes d'information 2007.
- [13] D'Hondt Frédéric & El Khayati Brahim : Etude des méthodes de clustering pour la segmentation d'images couleur. 2005.
- [14] Jiawei Han & Micheline Kamber : Data mining Concepts and technics second edition.
- [15] Christine Decaestecker et Marko Saerens : Analyse en composantes principales.
- [16] Nicolas Turenne : PCA Principal componenet analysis. INRIA cours sur les méthodes du data mining 2006.
- [17] Two cross company : Introduction to data mining and Knowledge discovery third edition.
- [18] Nicolas Turenne : Méthode des Knn. INRIA cours sur les méthodes du data mining 2006.
- [19] Prairie, James R., Rajagopalan, Balaji, Fulp, Terry J.; Zagona Modèle Modifié de K-NN pour la simulation stochastique d'écoulement de jet. Juillet 2006.



## Dynamic Load Balancing Strategy for Grid Computing

Belabbas Yagoubi<sup>1</sup> and Meriem Medebber<sup>2</sup>

<sup>1</sup> University of Oran, Faculty of Sciences, Department of Computer Science, Oran, Algeria

[byagoubi@yahoo.fr](mailto:byagoubi@yahoo.fr)

<sup>2</sup> University of Mascara, Department of Computer Science, Mascara, Algeria

[m.meddeber@yahoo.fr](mailto:m.meddeber@yahoo.fr)

**Abstract.** To improve the global throughput of the *Grid* infrastructures, effective and efficient load balancing algorithms are fundamentally important. Although load balancing problem in traditional parallel and distributed systems has been intensively studied, new challenges in Grid computing still make it an interesting topic, and many research projects are under way. This is due to the characteristics of Grid computing and the complex nature of the problem itself. In this paper, we propose a tree-based representation of a Grid, over which we develop a hierarchical load balancing strategy. The main characteristics of the proposed strategy can be summarized as follows: (i) It uses a task-level load balancing; (ii) It privileges local tasks transfer to reduce communication costs; and, (iii) It is a distributed strategy with local decision making.

**Keywords:** Load balancing; Grid computing; Tree-based model; Workload.

### 1 Introduction

Decrease in powerful computers costs coupled with advances in computer networking technologies have led to increased interest in the use of large-scale heterogeneous distributed systems. Indeed, recent researches on computing architectures are allowed the emergence of a new computing paradigm known as *Grid computing* [1].

In order to fulfil the user expectations in terms of performance and efficiency, the Grid system needs efficient load balancing algorithms for the distribution of tasks among their available resources. Hence, an important issue in such systems is the effective and efficient assignment of tasks, commonly referred to as *load balancing problem* and known to be as *NP-complete* problem [2].

The essential objective of a load balancing can be, depending on the user or the system administrator, defined by:

- The aim for the user is to minimize the makespan of its own application, regardless the performance of other applications in the system.
- The main goal for a system administrator is to maximize meet the tasks deadline by ensuring maximal utilization of available resources.

Typically, a load balancing scheme consists of four policies [3]:

1. *Information policy*: This policy is responsible to define when and how the information on the Grid resources availability is updated.
2. *Location Policy*: This policy is responsible to find a suitable transfer partner once that the transference policy decided that a given resource is server or receiver.
3. *Selection Policy*: This policy defines the task that should be transferred from the busiest resource to the idlest one.
4. *Transference Policy*: This policy classifies a resource as server or receiver of tasks according to its availability status.

Although load balancing problem in classical distributed systems has been intensively studied, new challenges in Grid computing still make it an interesting topic, and many research projects are under way. This is due to the specific characteristics of computational Grids [1,4], which make the load balancing problem more complex.

Our main contributions in this paper are two folds. We propose first a tree structure to represent any Grid architecture. Based on this tree representation, we develop a load balancing strategy which addresses both following objectives: (i) Reduction, as much as possible, of the average response time of tasks submitted to a Grid computing; and, (ii) Reduction of the communication costs.

The remainder of this paper is organized as follows. Some related works are described in Section 2. The mapping of any Grid architecture into a tree is explained in Section 3. Section 4 introduces the main steps of the proposed load balancing strategy. Associated load balancing algorithms are depicted in Section 5. Section 6 discusses behaviors of the proposed strategy through experimentation. Finally, the main conclusions of our research work are summarized and previews of future research are highlighted in Section 7.

## 2 Related works

Most application-level load balancing approaches are oriented on application partitioning via graph algorithms [5]. However, it does not address the issue of reducing migration cost, that is, the cost entailed by load redistribution, which can consume order of magnitude more time than the actual computation of a new decomposition. Some works [6] have proposed latency - tolerant algorithm that takes advantage of overlapping the computation of internal data and the communication of incoming data to reduce data migration cost. Unfortunately, it requires applications to provide such a parallelism between data processing and migration, which restricts its applicability. Payli et al. [7] proposed a dynamic load balancing approach to provide application level load balancing for individual parallel jobs in Grid computing environment. In [8], Hu et al. propose an optimal data migration algorithm in diffusive dynamic load balancing through the calculation of Lagrange multiplier of the Euclidean form of transferred weight. This work can effectively minimize the data movement in homogenous environments, but it does not consider the network heterogeneity.

The traditional objective, when scheduling sets of computational tasks, is to minimize the overall execution time called makespan. However, in the context of heterogeneous distributed platforms, makespan minimization problems are in most cases *NP-complete*, sometimes even *APX-complete* [2]. But, when dealing with large scale systems, an absolute minimization of the total execution time is not the only objective of a load balancing strategy. The communication cost, induced by load redistribution, is also a critical issue. For this purpose, we propose, in this paper, a novel load balancing strategy to address the new challenges in Grid computing. Comparatively to the existing works, the main characteristics of our strategy are:

(i) It uses a task-level load balancing; (ii) It privileges local tasks transfer than global, to reduce communication costs; and, (iii) It is a distributed strategy with local decision making.

## 3 Tree-based load balancing model

From topological point of view, we regard a Grid computing as a set of clusters in a multi-nodes platform. Each cluster owns a set of worker nodes and belongs to a local domain, i.e. a LAN (Local Area Network). Every cluster is connected to the global network or WAN (World Area Network).

The load balancing strategy proposed in this paper is based on mapping of any Grid architecture into a tree structure. This tree is build by aggregation as follows:

1. First, for each cluster we create a two levels tree. The leaves of this tree correspond to the nodes of the cluster, and its root, called *cluster manager*, represents a virtual tree node associated to the cluster, whose role is to manage the cluster workload.

2. Second, sub trees corresponding to all clusters are aggregated to build a three levels sub tree whose root is a virtual tree node designated as *Grid manager*.

The final tree is denoted by  $C/N$ , where  $C$  is the number of clusters that compose the Grid and  $N$  the total number of worker nodes in the Grid.

As illustrated in Figure 1, this tree can be transformed into two specific trees:  $C/N$  and  $I/N$ , depending on the values of  $C$  and  $N$ . The mapping function generates a non cyclic connected graph where each level has specific functions.

**Level 0:** In this first level, we have the *Grid manager*, which realizes the following:

- Maintains the workload information about the cluster managers.
- Decides to start a global load balancing between the clusters of the Grid, which we will call *intra-Grid load balancing*.
- Sends the load balancing decisions to the cluster managers for execution.

**Level 1:** Each *cluster manager* of this level is associated to a physical cluster of the Grid. In our load balancing strategy, this manager is responsible to:

- Maintain the workload information relating to each one of its worker nodes.
- Estimate the workload of associated cluster.
- Decide to start a local load balancing: *intra-cluster load balancing*.
- Send the load balancing decisions to the worker nodes which it manages, for execution.

**Level 2:** At this last level, we find the worker nodes of the Grid linked to their respective clusters. Each node at this level is responsible to:

- Maintain its workload information.
- Send, periodically, this information to its cluster manager.
- Perform the load balancing operations decided by its manager.

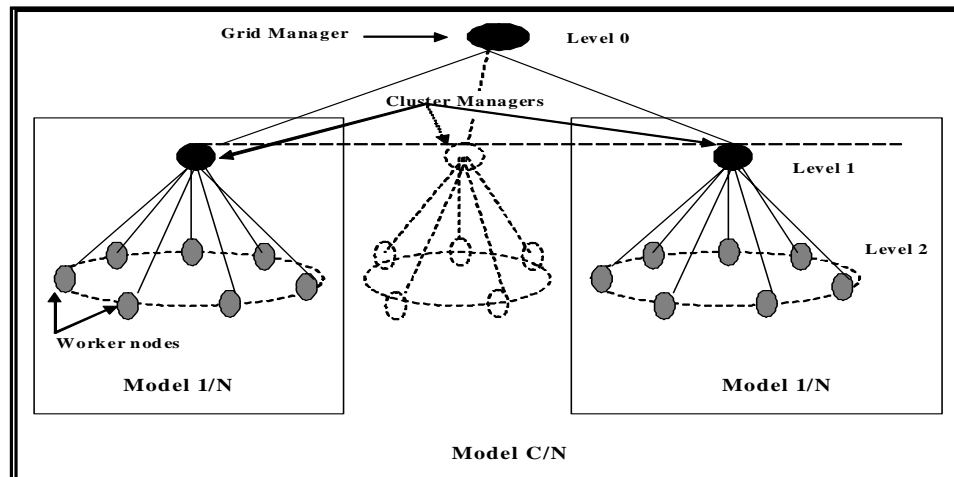


Fig. 1. Tree-based representation of a Grid

The proposed model is characterized by follows:

1. It is *hierarchical*: This propriety will facilitate the information flow through the tree.
2. It supports *heterogeneity* and *scalability* of Grids: Connecting/disconnecting Grid resources corresponds to adding/removing leaves or sub trees from the tree model.
3. It is totally *independent* from any physical architecture of a Grid: The transformation of a Grid topology into a tree is univocal. For each Grid, we associate one and only one tree, independently of the Grid topological complexity.

## 4 Load balancing strategy

### 4.1 Principles

We distinguish two load balancing levels: *Intra-cluster* and *Intra-Grid*:

- *Intra-cluster load balancing*: In this first level, depending on current workload of its associated cluster, estimated from its own worker nodes, each *cluster manager* decides whether to start or not a local load balancing operation.
- *Intra-Grid load balancing*: The load balancing at this level is performed only if some cluster managers fail to load balance their workload among their associated worker nodes.

The main advantage of this strategy is to privilege local load balancing in first (within a cluster). The goal of this neighbourhood approach is to decrease the amount of messages exchanged between clusters.

### 4.2 Strategy outline

At any load balancing level, we propose the following three steps strategy. As the description will be done in a generic way, we will use the concept of *group* and *element*. Depending on cases, a group designs either a cluster or the Grid. An element is a group component. The main steps of our strategy can be summarized as follows:

**Step 1- Estimation of the current group workload**

Knowing the number of available elements under its control and their computing capabilities, each group manager estimates its associated group capability by performing the following actions:

- Estimates current workload of the group based on workload information received from its component elements.
- Computes the standard deviation over the workload index in order to measure the deviations between its involved elements.
- Sends workload information to its Grid manager.

**Step 2- Decision making**

In this step the manager decides whether it is necessary to perform a load balancing operation or not. For this purpose it executes the two following actions:

1) *Defining the imbalance/saturation state of the group.* If we consider that the standard deviation measures the average deviation between the processing times of elements and f their group, we can say that this group is in balance state when this deviation is small. Then, we define the balance and saturation states.

- **Balance state:** We define a *balance threshold*, denoted as  $\epsilon$ , from which we can say that the standard deviation tends to zero and hence the group is balanced.
- **Saturation state:** A group can be balanced while being saturated. In this case, it is not useful to start an intra group load balancing since its elements will remain overloaded. To measure saturation, we introduce another threshold called *saturation threshold*, denoted as  $\delta$ .

2) *Group partitioning.* For an imbalance case, we determine the overloaded elements (*sources*) and the under loaded ones (*receivers*), depending on processing time of every element relatively to average processing time of the associated group.

**Step 3- Tasks transfer**

In order to transfer tasks from overloaded elements to under loaded ones, we propose the following heuristic:

- (i) Evaluate the total amount of load "Supply", available on receiver elements.
- (ii) Compute the total amount of load "Demand", required by source elements.
- (iii) If the supply is much lower than the demand (supply is far to satisfying the request) it is not recommended to start local load balancing. We introduce a third threshold, called *expectation threshold* denoted  $\rho$ , to measure relative deviation between supply and demand.
- (iv) Performs tasks transfer regarding communication cost induced by this transfer and according to selection criteria. As criterion selection we propose the *longest process time*, which transfer the task with longest remaining processing time.

**4.3 Supply and Demand estimation**

The supply of a receiver element  $Er$  corresponds to the amount of load  $Xr$  that it agrees to receive so that it's processing time:  $TEXr \in [TEX_G - \sigma_G; TEX_G + \sigma_G]$ ,

where  $\sigma_G$  corresponds to the standard deviation over the processing time of the group associated to element  $Er$ , and  $TEX_G$  represents the processing time of this group. In practice, we must reach the convergence:  $TEX_r \rightarrow TEX_G$

$$TEX_r = \frac{LODr + Xr}{SPDr} \cong \frac{LODG}{SPDG} \Rightarrow Xr = \frac{LODG \cdot SPDr}{SPDG} - LODr$$

Thus, we estimate the total supply of receivers set GER by:  $Supply = \sum_{Er \in GER} Xr$

By similar reasoning we determine the demand of a source element  $Es$ .

## 5 Generic load balancing algorithm

We define two levels of load balancing algorithms: *intra-cluster* and *intra-Grid*.

Each algorithm is triggered when the group manager notes that there is a load imbalance between the elements which are under its control. To do this report, the group manager receives, in a periodic way, workload information from each element.

**1 Intra cluster load balancing algorithm.** The neighborhood idea which privileges local load balancing in first makes that it is the most frequently requested level. It is executed when cluster managers find that imbalance exists between worker nodes under their control.

**2 Intra Grid load balancing algorithm.** It is started in the extreme case where majority of the cluster managers fail to locally balance their overload. Knowing the global state of each own cluster, the Grid manager can evenly distribute the global overload between its clusters. This load balancing level must, as far as possible, be avoided because inter clusters communication costs are very significant. This overhead is due primarily to the strong heterogeneity of the Grid resources.

### Step 1: Workload estimation

1. Workload information about each element  $E_i$  of  $G$ :  
**For** Every  $E_i$  **AND** according to its period **do** Sends its workload  $LOD_i$  to its group manager **End For**
2. According to its period the group manager performs:
  - a- Computes speed  $SPD_G$  and capacity  $SAT_G$  of  $G$ .
  - b- Evaluates current load  $LOD_G$  and processing time  $TEX_G$ .
  - c- Computes the standard deviation  $\sigma_G$  over processing times.
  - d- Sends workload information of  $G$  to its manager.

### ===== Step 2: Decision making =====

3. Balance criteria:  
**Switch**
  - $G = \text{cluster}$ : **If** ( $\sigma_G \leq \varepsilon$ ) **Then** Cluster in balance state; Return **End If**
  - $G = \text{Grid}$ : **If** (#overloaded clusters  $\leq$  Given threshold) **Then** Grid is balanced; Return **End If****End Switch**
4. Saturation criteria: **If** ( $LOD_G / SAT_G > \delta$ ) **Then** Group  $G$  is saturated; balancing Fail; Return **End If**
5. Partitioning  $G$  into overloaded ( $GES$ ), underloaded ( $GER$ ) and balanced ( $GEN$ ):  
 $GES \leftarrow \emptyset$ ;  $GER \leftarrow \emptyset$ ;  $GEN \leftarrow \emptyset$
- For** Every element  $E_i$  of  $G$  **do** **If** ( $E_i$  saturated) **Then**  $GES \leftarrow GES \cup \{E_i\}$
- Else Switch**
  - $TEX_i > TEX_G + \sigma_G \Rightarrow GES \leftarrow GES \cup \{E_i\}$
  - $TEX_i < TEX_G - \sigma_G \Rightarrow GER \leftarrow GER \cup \{E_i\}$
  - $TEX_G - \sigma_G \leq TEX_i \leq TEX_G + \sigma_G \Rightarrow GEN \leftarrow GEN \cup \{E_i\}$**End Switch**
- End If**
- End For**

### Step 3: Tasks transfer

$$6. \text{Supply} \equiv \sum_{E_r \in GER} \frac{LOD_r \cdot SPD_r}{SPD_G} - LOD_r \quad \text{Demand} \equiv \sum_{E_s \in GES} LOD_s - \frac{LOD_G \cdot SPD_s}{SPD_G}$$

**If** ( $\frac{\text{Supply}}{\text{Demand}} \leq \rho$ ) **Then** local Load balancing Fail; Return **End If**

7. Tasks transferring: **If** ( $G = \text{Cluster}$ ) **Then** Perform Heuristic1 **Else** Perform Heuristic2 **End If**

===== **Heuristic 1: Intra-cluster tasks transfer** =====

-Sort elements of  $GES$  by descending order of processing times.

-Sort elements of  $GER$  by ascending order of processing times.

**While** (  $GES \neq \phi$  **And**  $GER \neq \phi$  ) **do**

**For**  $i = 1$  **To** # (GER) **do**

        (i) Sort tasks of first node of  $GES$  by selection criterion,

        (ii) Transfer the higher priority task from first node of  $GES$  to  $i^{th}$  node of  $GER$ ,

        (iii) Update the current workloads of receiver and source elements,

        (iv) Update sets  $GES$ ,  $GER$  and  $GEN$ ,

        (v) **If** (  $GES = \phi$  **OR**  $GER = \phi$  ) **Then** Return **End If**

        (vi) Sort  $GES$  by descending order of their processing time

**End For**

**Done**

===== **Heuristic 2: Intra-Grid tasks transfer** =====

- Sort items of  $GES$  by descending order of their processing times.

**For** Every cluster  $C_j$  of set  $GES$  **do**

    (i) Sort clusters  $C_r$  of  $GER$  by ascending order of inter clusters ( $C_j - C_r$ ) WAN bandwidth sizes.

    (ii) Sort the nodes of  $C_j$  by descending order of their processing times.

    (iii) **While** (  $GES \neq \phi$  **And**  $GER \neq \phi$  ) **do**

**For**  $i = 1$  **To** # (GER) **do**

            (a) Sort tasks of first node belonging to  $C_j$  by selection criterion and communication cost,

            (b) Transfer the higher priority task from first node of  $C_j$  to  $i^{th}$  cluster of  $GER$ ,

            (c) Update the current workloads of source and receiver clusters,

            (d) Update sets  $GES$ ,  $GER$  and  $GEN$ ,

            (e) **If** (  $GES = \phi$  **OR**  $GER = \phi$  ) **then** Return **End If**

            (f) Sort  $GES$  by descending order of their processing times.

**End For**

**Done**

**End For**

## 6 Experimental Studies

### 6.1 Parameters

In order to evaluate the performance of our strategy we have implemented the associated algorithms on the GridSim simulator [9], which we extended to support simulation of varying Grid load balancing problems. We focused on the following objectives relating to a set of tasks submitted during a given period: *Average waiting time*, *Average execution time* and *Average response time*. To evaluate the benefit of our strategy, we compute the above metrics *before* (denoted  $\cdot$ ) and *after* ( $\cdot$ ) execution of our load balancing algorithm. All the experiments were performed on 3 GHz P4 Intel Pentium with 1 GB main memory, running on Linux Redhat 9.0.

### 6.2 Results

#### Experiments 1: Intra-cluster load balancing

In the first set of experiments we focused results relating to objective metrics, according to various numbers of tasks and worker nodes, in intra-cluster load balancing. We have varied the nodes number from 100 to 500 by step of 100. For each node we randomly generate associated speed varying between 5 and 40 MIPS. The number of tasks has been varied from 5000 to 10000 by step of 1000, with sizes randomly generated between 1000 and 200000 MI (Million of Instructions). Table 1 shows the variation of the average response time (in seconds) before and after execution of our load balancing strategy.

- Proposed strategy allowed reducing in a very clear way the mean response time of the tasks. We obtain a gain in more than 95% of cases, varying between 1% and 57%.
- The lower improvements were obtained when the number of nodes exceeds 300. We can justify this by the instability of the Grid state (nodes are largely under-loaded or even idle).

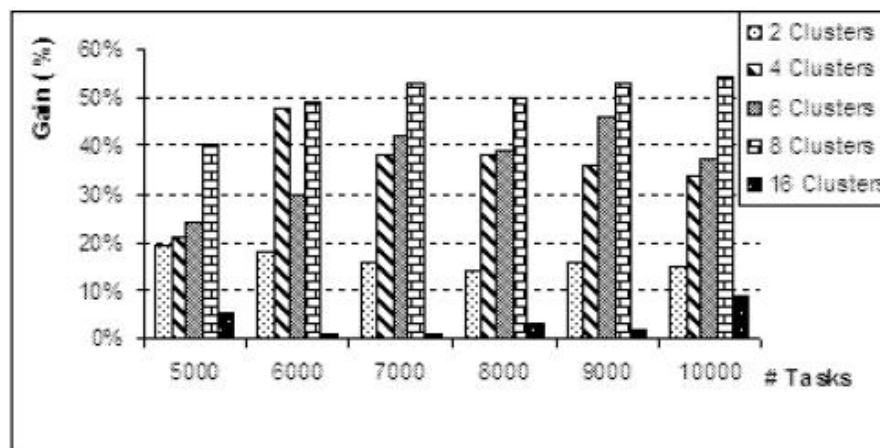
**Table 3.** Improvement Realized By Intra-Cluster Strategy.

#Tasks	#Nodes	100	200	300	400	500
<b>5000</b>	<i>Bef</i>	4.78E+05	2.32E+04	4 4.92E+03	1.75E+03	1.29E+03
	<i>Aft</i>	3.32E+05	1.19E+04	3.48E+03	1.64E+03	1.28E+03
	<i>Gain</i>	<b>31%</b>	<b>49%</b>	<b>29%</b>	<b>6%</b>	<b>1%</b>
<b>6000</b>	<i>Bef</i>	6.00E+05	3.03E+04	6.44E+03	2.28E+03	1.64E+03
	<i>Aft</i>	4.21E+05	1.49E+04	3.81E+03	2.11E+03	1.64E+03
	<i>Gain</i>	<b>30%</b>	<b>51%</b>	<b>41%</b>	<b>7%</b>	<b>0%</b>
<b>7000</b>	<i>Bef</i>	7.43E+05	3.89E+04	8.31E+03	2.89E+03	2.07E+03
	<i>Aft</i>	5.21E+05	1.94E+04	4.84E+03	2.67E+03	2.07E+03
	<i>Gain</i>	<b>30%</b>	<b>50%</b>	<b>42%</b>	<b>8%</b>	<b>0%</b>
<b>8000</b>	<i>Bef</i>	9.06E+05	4.86E+04	1.05E+04	3.63E+03	2.59E+03
	<i>Aft</i>	6.46E+05	2.43E+04	5.76E+03	3.04E+03	2.54E+03
	<i>Gain</i>	<b>29%</b>	<b>50%</b>	<b>45%</b>	<b>16%</b>	<b>2%</b>
<b>9000</b>	<i>Bef</i>	1.09E+06	6.04E+04	1.32E+04	4.56E+03	3.22E+03
	<i>Aft</i>	7.84E+05	3.00E+04	6.69E+03	3.84E+03	3.11E+03
	<i>Gain</i>	<b>28%</b>	<b>50%</b>	<b>49%</b>	<b>16%</b>	<b>3%</b>
<b>10000</b>	<i>Bef</i>	1.30E+06	7.42E+04	1.63E+04	5.63E+03	3.96E+03
	<i>Aft</i>	9.39E+05	3.93E+04	7.03E+03	4.30E+03	3.61E+03
	<i>Gain</i>	<b>28%</b>	<b>47%</b>	<b>57%</b>	<b>24%</b>	<b>9%</b>

### Experiments 2: Intra-Grid load balancing

During these experiments, we interested to the intra-Grid load balancing behaviours. We have considered different numbers of clusters and we supposed that each cluster involves 30 worker nodes. Figure 2 illustrates the improvement of the mean response time, obtained by our load balancing strategy for various numbers of clusters by varying the number of tasks.

- Except the case of 16 clusters, all the profits are higher than 10%. We consider this important behaviour very promising.
- The lower benefits were obtained when the number of clusters was set to 16. We can justify this by the instability of the Grid state (Most nodes are under loaded or even idle).

**Fig. 2.** Gain according to various numbers of clusters.

## 7 Conclusion and future works

This paper addressed the problem of load balancing in Grid computing. We proposed a load balancing model based on a tree representation of a Grid. The model takes into account the heterogeneity of the resources and it is completely independent from any Grid physical architecture. Basing on this model, we defined a hierarchical load balancing strategy having two main objectives: (i) Reduction of the mean response time of tasks submitted to a Grid computing; and, (ii) Reduction of the communication costs during task transferring.

Relatively to the existing works, our strategy uses a task-level load balancing and privileges, as much as possible, a local load balancing to avoid the use of WAN communication. It is a distributed strategy with local decision-making. Taking into account the over cost generated by the execution of load balancing system, our strategy starts a balancing operation only when it ensures that it is convenient.

The first results obtained on *GridSim* simulator are very promising. We have appreciably improved the average response time.

In the future, we plan to integrate our load balancing strategy on others known Grid simulators like *SimGrid* [10]. This will allow us to measure the effectiveness of our strategy in existing simulators. As another perspective, we think to extend the proposed model to a fully distributed model (removal of the root from the tree structure). We also envisage developing our strategy as a service of GLOBUS middleware [11].

## References

- [1] I. Foster and C. Kesselman (editors). 'The Grid2: Blueprint for a New Computing Infrastructure', Morgan Kaufmann (second edition), USA, (2004).
- [2] C.Z. Xu and F.C.M. Lau. 'Load Balancing in Parallel Computers: Theory and Practice'. Kluwer, Boston, MA, (1997).
- [3] K. Kurowski, B. Ludwiczak, J. Nabrzyski, A. Oleksiak, and Pukacki. Improving grid level throughput using job migration and rescheduling. *Scientific Programming*, vol.12 number 4, (2004), pp 263-273.
- [4] M. Baker, R. Buyya, and D. Laforenza. Grids and grid technologies for wide-area distributed computing. *Journal of Software Practice and Experience*, Vol.32 number 15, (2002), pp 1437-1466.
- [5] M. T. Kechadi and K. Savvas. Dynamic task scheduling for irregular network topologies. *Parallel Computing 31, Elsevier journal*, (2005), pp 757-776.
- [6] H. Shan, L. Oliker, R. Biswas, and W. Smith. Scheduling in heterogeneous grid environments: The effects of data migration. In *Proceedings of ADCOM2004: International Conference on Advanced Computing and Communication*, India, (2004).
- [7] R.U. Payli, E. Yilmaz, A. Ecer, H.U. Akay, and S. Chien. DLB: A dynamic load balancing tool for grid computing. *Parallel CFD Conference*, Grand Canaria, Canary Islands, Spain, (2004), pp 24-27.
- [8] Y.F. Hu, R.J. Blake, and D.R. Emerson. An optimal migration algorithm for dynamic load balancing. *Concurrency: Practice and Experience*, vol.10, (1998), pp 467-483.
- [9] R. Buyya. A grid simulation toolkit for resource modelling and application scheduling for parallel and distributed computing.. <http://www.buyya.com/gridsim/>
- [10] H. Casanova, Simgrid: A Toolkit for the Simulation of Application Scheduling, *Proceedings of the First IEEE/ACM Int. Symposium on Cluster Computing and the Grid*, Brisbane, Australia, (2001), pp430-437.
- [11] I. Foster. Globus toolkit version 4: Software for service oriented systems. *International Conference on Network and Parallel Computing*. Beijing, China, (2005), pp2-13.



## Extension conceptuelle de méta-modèles XML et UML pour une transformation bidirectionnelle de modèles

Ludovic Menet\*, \*\*, Myriam Lamolle\*

\* Laboratoire d'Informatique et Communication (LINC)  
IUT de Montreuil, Université Paris8  
140 rue de la Nouvelle France, F-93100 Montreuil  
{m.lamolle, l.menet}@iut.univ-paris8.fr

\*\* Orchestra Networks, Direction des Etudes et Recherches  
75 boulevard Haussman, 75008 Paris  
ludovic.menet@orchestranetworks.com

**Résumé.** XML est devenu depuis la fin des années 90 le standard pour échanger et envoyer de l'information sur Internet. Le W3C a ensuite préconisé l'utilisation de XML Schema pour définir la structure des documents XML. A ce jour, la modélisation graphique de modèles XML Schema n'est pas standardisée. L'introduction d'un formalisme de définition de modèles est un moyen de rendre la modélisation plus accessible. UML est un langage de modélisation objet de plus en plus utilisé et reconnu aujourd'hui comme un standard dans le domaine du génie logiciel, ce qui en fait un candidat idéal pour la modélisation des modèles XML Schema. Dans cet article, nous introduisons les spécificités du formalisme UML afin de faciliter la définition des modèles XML. Un enrichissement conceptuel est réalisé dans UML et XML Schema dans le but de réaliser un mapping bidirectionnel entre ces deux standards.

**Mots-clés:** UML, Profil UML, XML, XML Schema, Mapping, Meta-modèle.

### 1 Introduction

Il est désormais admis que le langage XML [18] est devenu le standard pour échanger de l'information au travers d'Internet. XML Schema [19] est un formalisme permettant de décrire la structure d'un document XML de façon beaucoup plus précise qu'une simple DTD (Document Type Definition). L'utilisation d'XML s'avère adaptée à la définition de modèles mais implique une connaissance étendue de ce langage. De nombreux logiciels tels que Altova XML Spy et oXygen XML Editor ont été développés dans le but de modéliser graphiquement, sous la forme d'arbres, des modèles XML Schema. Ces logiciels permettent d'optimiser la modélisation des schémas XML mais proposent un formalisme de représentation différent d'une solution à l'autre. Il découle de ce constat qu'un formalisme standard devrait être utilisé afin de faciliter la modélisation et le transfert de ces modèles. Cet article a pour objectif de présenter des techniques de modélisation de modèles XML Schema par l'intermédiaire des diagrammes de classes appartenant au standard UML (Unified Modelling language).

UML propose un langage de modélisation graphique très abouti et grandement utilisé en génie logiciel. Notre objectif est d'introduire toute la richesse du formalisme UML pour faciliter la modélisation des modèles XML Schema. L'introduction du formalisme UML dans notre architecture XML est réalisée par un mapping entre les méta-modèles de ces deux standards. Pour ce faire, nous procédons en plusieurs étapes à savoir :

1. définition d'un méta-modèle décrivant les concepts supportés par XML Schema ;
2. définition d'extensions XML Schema afin de prendre en compte toutes les spécificités objets d'UML.  
Ces extensions exploitent les mécanismes d'extension proposés par XML Schema afin de rester conforme à la norme du W3C ;
3. définition d'un profil UML afin de spécialiser UML à la sémantique des modèles XML Schema ;
4. définition des mappings entre les méta-modèles d'UML et XML Schema ;
5. application des mappings par l'intermédiaire du format pivot *XMI* et transformation des modèles par feuilles de styles *XSLT* décrivant les mappings.

## 2 Etats de l'art

Levendovszky [11], se basant sur les travaux de Lemesle [10], définit le mapping comme étant un ensemble de règles de transformation de modèles permettant de traduire des instances d'un méta-modèle source en instance d'un méta-modèle cible.

Baïna et al. [3] apportent une définition mathématique à l'interopérabilité des applications dans un système d'entreprise. Après étude de ces travaux, nous avons conclu que cette définition peut aussi s'appliquer aux mapping de modèles. Considérons  $A$  et  $B$  deux méta-modèles ;  $A$  et  $B$  sont dits interopérables si et seulement s'il existe un mapping bijectif de  $M_A$  vers  $M_B$ , que nous noterons  $f$ . La bijection de  $f$  nous assure que l'on peut construire une instance de modèle  $B$  à partir de l'instanciation du modèle de  $A$  (en utilisant  $f$ ) et construire une instance de modèle  $A$  à partir de l'instanciation du modèle de  $B$ . A partir de cette définition, trois niveaux d'interopérabilité sont définis et des mappings sont identifiés entre les langages  $A$  et  $B$  à savoir :

- Niveau 2 : Il existe un isomorphisme total entre  $M_A$  et  $M_B$ . De ce fait, tout concept de  $M_A$  a son équivalent dans  $M_B$  et inversement, ce qui signifie que  $M_A$  et  $M_B$  sont équivalents ;
- Niveau 1 : Il existe un isomorphisme partiel entre  $M_A$  et  $M_B$ . Il existe donc une sous-partie de  $M_A$  que l'on notera  $M'_A$  et une sous-partie de  $M_B$  ( $M'_B$ ) telles que l'interopérabilité entre  $M'_A$  et  $M'_B$  est de niveau 2, ces sous-parties sont donc équivalentes ;
- Niveau 0 : Il n'existe pas d'isomorphisme partiel entre  $M_A$  et  $M_B$ . Cependant il se peut que des mappings non bijectifs existent entre  $M_A$  et  $M_B$  ; dans ce cas, nous ne pouvons pas parler d'interopérabilité sémantique en  $A$  et  $B$ .

L'interopérabilité de niveau 2 est la plus difficile à établir étant donné qu'il est très rare que deux méta-modèles soient totalement équivalents. Dans cette étude, nous allons donc chercher à établir une correspondance partielle entre  $M_A$  et  $M_B$  en calculant la proportion de  $M'_A$  et  $M'_B$  par rapport à  $M_A$  et  $M_B$  obtenant ainsi une interopérabilité de niveau 1 entre XML Schema et UML.

Carlson [4] aborde la transformation de modèles UML en schéma XML et l'opération inverse, obtenant ainsi un mapping bijectif entre ces deux formalismes. Le mapping réalisé utilise comme fondement un profil UML définissant des concepts spécifiques à XML Schema. Ce profil est utilisé afin d'étendre la sémantique d'un modèle UML à la sémantique d'un modèle XML Schema. Cette approche permet un mapping d'une grande partie des concepts introduits par XML Schema, mais ne prend pas en compte certains concepts tels que la notion de groupes (list, union), de contrainte d'identité (key, keyref, unique), etc. De plus, certains concepts importants d'UML, tels que l'agrégation, la composition, l'association et la documentation, ne sont pas pris en compte lors d'une transformation d'un modèle UML vers un modèle XML Schema. Ce mapping a été mis en application par *HyperModel* [5]. *HyperModel* est un plug-in pour l'IDE Eclipse implémentant le mapping bijectif UML/XML. Cet outil est fonctionnellement opérationnel mais souffre de quelques limitations lors d'une transformation d'un modèle UML vers un modèle XML Schema car :

1. certains concepts ne sont pas mappés (agrégation, composition, etc.) ;
2. des éléments sont mappés plusieurs fois entraînant une redondance d'information et une inconsistance dans le modèle résultat ;
3. une perte d'information, plus précisément concernant les contraintes de cardinalité, se produit sur certains modèles ;
4. des modèles XML Schema générés ne sont pas valides au regard de la spécification du W3C.

Routledge et al. [14] abordent le mapping de manière traditionnelle entre UML et XML Schema par l'intermédiaire de l'approche à trois niveaux issus du monde des bases de données à savoir les niveaux conceptuel, logique et physique. Dans le contexte d'un diagramme de classes UML, le niveau conceptuel décrit les objets et leurs relations. Le niveau logique représente les structures de données XML Schema sous la forme d'un profil UML. Le niveau physique représente directement le modèle XML Schema. De la même manière que les travaux de Carlson, certains éléments spécifiques d'UML tels que l'agrégation, la composition et d'autres ne sont pas pris en compte.

D'autres travaux ont été réalisés dans le même contexte par Conrad et al. [6], Kurtev et al. [9] mais souffrent également des mêmes limitations que les travaux que nous avons présentés précédemment.

### 3 Enrichissement conceptuel de XML Schema et d'UML

A ce jour, la modélisation graphique de modèles XML Schema n'est pas standardisée. Il existe certes des outils pour la modélisation de modèles XML mais ceux-ci sont restreints à la sémantique d'XML Schema (Altova XML Spy [1], oXygen XML editor [17], etc.). En effet, ces outils sont dans l'incapacité de guider l'utilisateur dans l'utilisation des concepts introduits par des extensions spécifiques. L'introduction d'un formalisme de définition de modèle est un moyen de rendre la modélisation plus accessible. UML est un langage de modélisation objet de plus en plus utilisé et reconnu aujourd'hui comme un standard dans le domaine du génie logiciel, ce qui en fait un candidat idéal pour la modélisation des modèles XML Schema. Dans l'objectif de faciliter et d'optimiser le mapping entre XML Schema et UML, il est nécessaire, dans un premier temps, d'enrichir chacun de ces formalismes avec les spécificités issues de l'autre. C'est une première étape d'homogénéisation des différents formalismes.

#### 3.1 Adjonction de métaconnaissances objet dans les modèles XML

UML est un formalisme de modélisation objet qui définit des notions telles que la généralisation, la composition et l'agrégation. Bien qu'il soit possible de matérialiser ces notions de manière implicite dans un modèle XML schéma, nous proposons d'introduire des métadonnées matérialisant ces notions de manière explicite dans les modèles XML Schema (notées *nom\_concept\_objet\_UML* à la ligne 3 de la figure 1). Pour ceci faire, nous utilisons les mécanismes d'extensions préconisés par XML Schema, soit pour chaque métaconnaissance, une description sous la forme de l'extension suivante :

```
1 ...<xs:annotation>
2   <xs:appinfo>
3     <osd:nom_concept_objet_UML/>
4   </xs:appinfo>
5 </xs:annotation>...
```

Fig. 1. Extension XML Schema représentant une métaconnaissance objet.

L'ajout de ces métadonnées dans les modèles XML Schema permet d'inclure des spécificités objet d'UML et de mettre en évidence des relations entre certains concepts. Au delà de notre processus de mapping, ces métadonnées contribuent à optimiser certains traitements tels que la factorisation de données, l'optimisation d'arbres, la suppression d'instances devenues inutiles [12].

#### 3.2 Spécialisation d'UML en modèle XML Schema

En plus de ses capacités de modélisation, UML permet de définir des profils. Un profil spécialise le formalisme UML pour un domaine ou une technologie particulière. De nombreux profils ont été développés, par exemple les profils CORBA [16] et EJB [7]. La possibilité de définir des profils UML a donné lieu à différents travaux tels que ceux de Mahmoud [13] et Ambler [2].

La définition d'un profil nous permet de spécialiser UML à la sémantique des modèles XML Schema. Le processus adéquat pour définir un profil UML est le suivant :

1. Définition du domaine du profil, soit le méta-modèle définissant les concepts et les relations que l'on souhaite introduire ;
2. Définition technique du profil consistant à établir une correspondance entre les concepts UML et ceux du profil ;
3. Définition d'un exemple simple et concret du profil.

La première étape est réalisée en exploitant le méta-modèle de XML Schema.

Le tableau 1 présente un extrait de la définition de notre profil UML représentant ce méta-modèle :

**Table 1.** Extrait de la définition du profil UML.

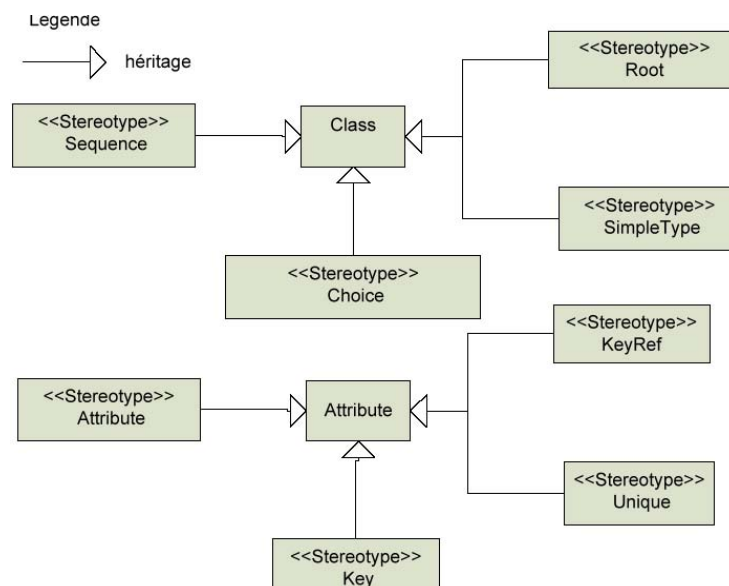
Stéréotype	Appliqué à	Description
Root	Classe	Définit la racine d'un modèle XML Schema
Sequence	Classe	Définit un élément complexe XML Schema de type 'sequence'
Choice	Classe	Définit un élément complexe XML Schema de type 'choice'
SimpleType	Classe	Définit un type redéfini au sens XML Schema
Attribute	Attribut	Définit un attribut au sens XML Schema

Le mécanisme d'extension d'UML nous permet d'étendre son formalisme à notre sémantique. Cette extension est réalisée par des *stéréotypes* et des *valeurs marquées*. Les stéréotypes sont utilisés pour définir un nouveau type d'élément à partir d'un élément existant du méta-modèle UML.

La figure 2 présente un extrait de l'application de cette définition de notre profil UML pour des modèles XML Schema.

Nous pouvons constater que les stéréotypes de la figure 2 (noté `<<Stereotype>>` *Root* par exemple) héritent de l'élément *Class* ou *Attribute* du méta-modèle UML. Par conséquent, chacun de ces stéréotypes sera instancié par le constructeur du méta-modèle de la même manière que l'élément *Class* ou *Attribute*. Des valeurs marquées peuvent, de plus, être associées à des stéréotypes. Celles-ci spécifient des paires *clés-valeurs* pour fixer un ensemble de propriétés d'éléments existants ou de stéréotypes définis. La définition de ces stéréotypes permet d'introduire plus de sémantique, extérieure à UML, qui nous permettra de représenter un modèle XML Schema à l'aide de diagrammes UML.

Dans cette section, nous avons enrichi conceptuellement XML Schema et UML dans le but d'établir des correspondances entre ces deux technologies. Ces correspondances nous permettent de spécifier des mappings entre UML et XML Schema.

**Fig. 2.** Extrait du profil UML représentant les spécificités de XML Schema.

## 4 Spécification des mappings UML / XML Schema

Nous présentons maintenant les règles de mapping entre éléments UML et XML Schema, y compris les notions décrivant des relations de dépendance.

### 4.1 Classes UML

Une classe UML est mappée vers un élément XML (*<xs:element>*) ayant une structure complexe ou vers un élément de type complexe global au schéma XML (*<xs:complexType>*). Certains stéréotypes que nous avons définis dans notre profil permettent d'ajouter des informations supplémentaires sur les classes. Ainsi, les stéréotypes *<<Sequence>>*, *<<Choice>>* et *<<All>>*, par exemple, permettent-ils d'indiquer que la classe doit être transformée en un élément complexe XML ayant respectivement une structure de la forme *<xs:sequence>*, *<xs:choice>* ou *<xs:all>*. La notion d'élément racine a été définie dans la sémantique de XML Schema sous la forme du stéréotype *<<Root>>*.

### 4.2 Attributs

Les attributs UML sont transformés en éléments simples XML *<xs:element>* ayant pour valeur d'attribut *<xs:type>* un type de donnée. L'élément XML généré est englobé soit par un élément *<xs:sequence>*, *<xs:choice>* ou *<xs:all>* selon le stéréotype appliqué à la classe. Nous avons introduit le stéréotype *<<Attribute>>* afin de spécifier que l'attribut UML doit être transformé en attribut XML *<xs:attribute>* et non en élément *<xs:element>*.

### 4.3 Cardinalités

Les cardinalités UML sont transformées en attributs *<xs:minOccurs>* et *<xs:maxOccurs>* portés par l'élément correspondant. Les cardinalités peuvent porter sur les associations, les agrégations, les compositions et les attributs.

### 4.4 Types de données natifs

Les types de données natifs UML tels que *int*, *double*, *float*, *string*, etc., sont mappés en entités XML correspondantes, soit respectivement *<xs:int>*, *<xs:double>*, *<xs:float>*, *<xs:string>*, etc.

### 4.5 Types dérivés

Un type dérivé permet de spécifier des contraintes sur des types ou de définir de nouveaux types métier. Nous définissons le stéréotype *<<SimpleType>>* permettant d'indiquer que l'élément UML correspondant est un type dérivé. Les propriétés du type dérivé sont définies dans une annotation. L'élément XML correspondant est construit à partir des informations définies dans l'annotation associée à l'élément UML. Dans notre mapping, nous incluons les contraintes XML sur les types. Il est ainsi possible pour un type dérivé de spécifier des contraintes conformes à la spécification du W3C telles que *length*, *min/max length*, *enumeration*, *fractionDigits*, *totalDigits*, *max/min Inclusive*, *max/min Exclusive*.

### 4.6 Généralisation

Une classe UML stéréotypée *<<abstract>>* est déclarée comme étant abstraite et étant une généralisation d'un ou plusieurs concepts (classes). Nous avons défini précédemment des extensions XML permettant de matérialiser des relations « objet ». Une généralisation sera donc traduite en XML par un élément possédant l'extension suivante :

```
<xs:annotation>
  <xs:appinfo>
    <osd:generalisation>
      <osd:conceptPath>
        Path du sous concept dans le schéma
      </osd:conceptPath>
    </osd:generalisation >
  </xs:appinfo>
</xs:annotation>
```

#### 4.7 Agrégation et composition

L'agrégation (respectivement composition) est traduite en XML par un élément possédant l'extension suivante :

```
<xs:annotation>
  <xs:appinfo>
    <osd:aggregation (respectivement composition) />
  </xs:appinfo>
</xs:annotation>
```

Les cardinalités UML définies pour les associations, agrégation et composition sont représentées par les attributs XML `<xs:minOccurs>` et `<xs:maxOccurs>`. De plus, les extensions définies pour ces mêmes relations sont définies uniquement dans l'élément XML parent de l'association.

#### 4.8 Documentations et notes

Une documentation ou une note est une variante affaiblie de contrainte ; c'est un texte à l'usage de l'utilisateur mais non exécutable par la machine. Dans un diagramme de classe UML, une note est rattachée à l'élément auquel elle fait référence. Les notes et documentations sont traduites en XML par l'extension suivante :

```
<xs:annotation>      <xs:documentation xml:lang = « fr FR »>
<osd:description> Texte </osd:description>      </xs:documentation>
</xs:annotation>
```

L'attribut `xml:lang` permet de spécifier à quelle langue est associée la note ou la documentation.

### 5 Application des mappings

Nous avons présenté dans la section précédente une spécification des mappings des éléments de base d'UML. UML est un formalisme de représentation graphique de modèles et XML Schema est un langage de description de modèles structurés. L'application des mappings nécessite de pouvoir représenter les modèles XML Schema et UML dans un formalisme commun.

#### 5.1 XML Metadata Interchange Format (XMI)

XMI [8] est un formalisme de représentation de modèle recommandé par l'ODMG [15] créé dans le but de pouvoir échanger des modèles entre outils de modélisation. La représentation UML d'un modèle représente le niveau abstrait de celui-ci, le niveau concret est matérialisé par la représentation XMI de ce modèle. XMI représente un format pivot permettant de représenter et manipuler un modèle UML sous la forme d'un document XML. Le Document XMI contient deux parties importantes à savoir un « *header* » et un « *content* ». La partie « *header* » définit les informations sur le modèle représenté (propriétaire, description, etc.). La partie « *content* » définit le contenu du modèle, c'est-à-dire les classes, les stéréotypes, les associations, etc. De nombreux logiciels de modélisation (comme ArgoUML, Rational Rose, Objecteering, etc.) incluent les fonctionnalités d'import et d'export de modèles UML au format XMI. Cependant, ces logiciels peuvent pour un même modèle générer des documents XMI valides mais ayant des structures différentes, rendant l'échange de modèles difficile entre ces différents outils. Cet inconvénient majeur nous oblige à appliquer nos propres mappings afin de générer des documents XML Schema, conformes aux normes du W3C.

#### 5.2 XSL Transformation

XMI permet de représenter un modèle UML sous la forme d'un document XML, permettant ainsi d'établir une corrélation entre un modèle UML en modèle XML Schema. XSLT, quant à lui, est un langage permettant d'appliquer des règles de transformation sur un document XML pour obtenir un autre document XML. Ces règles de transformation sont décrites dans des feuilles de styles XSL. La figure 3, analogue aux principes de Lemesle [10], décrit le processus de transformation :

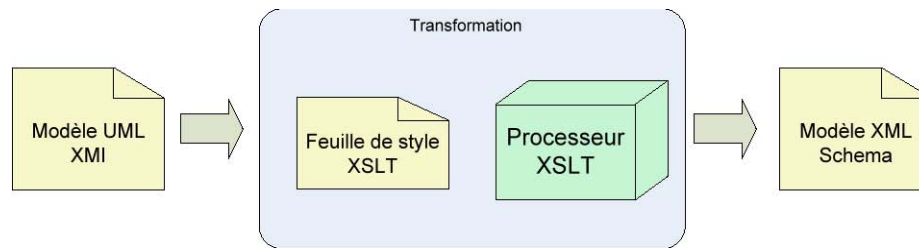


Fig. 3. Processus de transformation XSLT.

### 6.3 Etude de cas

Cette partie illustre un extrait de modèle UML et son équivalent XML Schema après application des règles de mappings que nous avons établies. La figure 4 présente la modélisation UML d'un réseau simplifié de trains. Ce diagramme est composé de différents concepts et relations entre ces derniers. Il est question dans notre exemple de mettre en avant les relations d'association, de composition, d'agrégation et de type dérivé. Ainsi avons nous par exemple défini le concept de *train* comme étant composé d'un *moteur*, de *roues* (notion de composition), pouvant posséder des *wagons* (notion d'agrégation) et ayant des propriétés telles qu'un *type* et une *marque*. Nous associons un conducteur de type *Personne* à un train. Ce concept porte des propriétés telles que *nom*, *prénom* et *date de naissance* typées par un type de données de base UML. Nous avons défini une propriété *email* représentant l'utilisation d'un type de données redéfini. La classe *Email* porte le stéréotype `<<SimpleType>>` permettant d'indiquer qu'il s'agit d'un type redéfini au sens XML Schema. Les propriétés de ce type redéfini sont contenues dans une annotation UML spécifiant des valeurs pour `SimpleType::base`, `SimpleType::pattern` et `SimpleType::whiteSpace`. La racine du schéma est matérialisée par le stéréotype `<<Root>>`, appliqué à la classe *ReseauTrains*. Les classes de ce modèle, excepté la classe *email*, sont stéréotypées `<<Sequence>>`. Ceci permet de spécifier que la classe correspondante représente un élément complexe, au sens XML Schema, de type *sequence*. A partir de ce diagramme, un document XMI est généré. Une feuille de style XSL est appliquée sur le document XMI obtenu. Cette feuille de style est composée de règles définissant les mappings que nous avons présentés en section 5.1

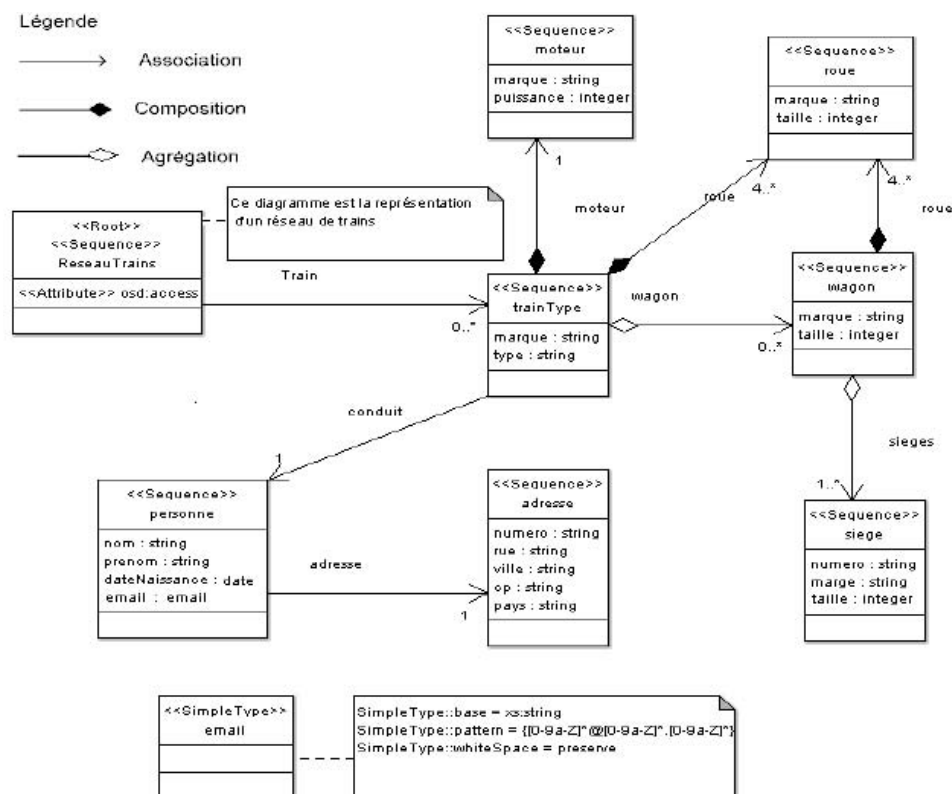


Fig. 4. Exemple de modèle UML.

La figure 5 présente la transformation de la classe *trainType* et de ses relations, en élément XML Schema :

```
<xs:complexType name="trainType">
```

```
  <xs:sequence>
```

1 Nous avons choisi de ne pas présenter notre feuille de style dans la mesure où elle ne relève que d'une implémentation technique de nos mappings.

```
    <xs:element name="marque" type="xs:string"/>
    <xs:element name="type" type="xs:string"/>
    <xs:element name="wagon" type="wagon" minOccurs="0"
      maxOccurs="unbounded">
      <xs:annotation><xs:appinfo>
        <osd:aggregation/>
      </xs:appinfo></xs:annotation>
    </xs:element>
    <xs:element name="moteur" type="moteur" minOccurs="1"
      maxOccurs="1">
      <xs:annotation><xs:appinfo>
        <osd:composition/>
      </xs:appinfo></xs:annotation>
    </xs:element>
    <xs:element name="roue" type="roue" minOccurs="4"
      maxOccurs="unbounded">
      <xs:annotation><xs:appinfo>
        <osd:composition/>
      </xs:appinfo></xs:annotation>
    </xs:element>
    <xs:element name="conduit" type="personne" minOccurs="1"
      maxOccurs="1"/>
  </xs:sequence>
</xs:complexType>
```

Fig. 5. Extrait d'un schéma XML généré à partir d'un modèle UML.

## 7 Conclusion

Nous avons montré dans cet article que nous pouvons utiliser le formalisme UML pour définir des modèles, facilitant ainsi la modélisation de ces derniers. L'introduction d'UML a été réalisée par la création de correspondances entre XML Schéma et UML, correspondances qui ont permis de définir des mappings entre ces deux standards. En introduisant des métaconnaissances dans un modèle XML Schéma et en définissant un profil UML, nous obtenons une interopérabilité de niveau 1 entre XML et UML, comblant certaines limitations que nous avons présentées dans la section 3. En effet, les extensions XML Schéma que nous avons introduites nous permettent d'établir des correspondances entre des concepts qui n'étaient pas pris en compte tels que les notions d'agrégation, de composition, documentation etc. De plus, l'ajout de ces notions, plus particulièrement les notions d'agrégation et de composition, permet de mettre explicitement en évidence des liens sémantiques entre éléments et d'apporter des informations concernant le cycle de vie de ces derniers.

La suite de nos travaux portera sur l'amélioration de nos mappings UML notamment en prenant en compte la définition de contraintes OCL (Object Constraint Language) [20] et le développement d'un outil permettant de définir parallèlement des modèles UML et XML Schéma afin de faciliter la semi automatisé des mappings et la validation des modèles par les experts du domaine.



## Références

1. Altova XMLSpy. <http://www.altova.com/xmlspy>
2. Ambler S. *Persistence Modeling in the UML*. Published in issue of Software Development. August. <http://www.sdmagazine.com> (1999)
3. Baïna S., Panetto H., Benali K.. Apport de l'approche MDA pour une interopérabilité sémantique. Ingénierie des Systèmes d'Information (ISI). Volume 11. p. 11-29. (2006)
4. Carlson D. *Modeling XML Applications with UML: Practical e-Business Applications*. Addison-Wesley Inc. (2001)
5. Carlson D. Semantic Models for XML Schema with UML Tooling, In Proceedings of the 2nd International Workshop on Semantic Web Enabled Software Engineering. (2006)
6. Conrad R., Scheffner, D. and Freytag, J.C. *XML conceptual modeling using UML*. Proceedings of the 19th International Conference on Conceptual Modeling (ER'2000).
7. Greenfield J. UML Profile For EJB. Rational Software Corp, May. (2001)
8. Iyengar, S. and S.A. Brodsky. *XML Metadata Interchange (XMI)*. Proposal to the OMG Object Analysis & Design Task Force RFP 3: Stream-based Model Interchange Format (SMIF), Object Management Group. <http://www.omg.org>. (1998).
9. Kurtev, I. *UML to XML-Schema Transformation: a Case Study in Managing Alternative Model Transformations in MDA*. Forum on specification and Design Languages. (2003)
10. Lemesle, R. *Transformation Rules Based on Meta-Modelling*. EDOC'98, La Jolla, California, 3-5 November 1998, p. 113-122. (1998)
11. Levendovszky T., Karsai G., Maroti M., Ledeczi A. and Charaf H. *Model Reuse with Metamodel-Based Transformations*. In Proceedings of the 7th International Conference on Software Reuse: Methods, Techniques, and Tools table of contents. p. 166-178. (2002)
12. Menet L., Lamolle M. *Meta-modelling "object": expression of semantic constraints in complex data structure*. Proceedings of ERIMA07'. p. 104-108. Biarritz, FRANCE. (2007)
13. Mahmoud N. *Analyse/conception par points de vue : le profil VUML*. Thèse de doctorat d'informatique, INSA Toulouse. (2005)
14. Routledge, N., Bird, L. and Goodchild, A. *UML and XML schema*. In Proceedings of the 13th Australasian Database Conference. p. 157-166. Melbourne, Australie. (2002)
15. ODMG. *The Object Data Standard : ODMG 3.0*. Morgan Kauffman Publishers. (1999)
16. OMG. "CORBA specifications". <http://www.omg.org/cgi-bin/doc?formal/02-04-01> (2002)
17. oXygen XML editor. <http://www.oxygenxml.com>
18. W3C. Extensible Markup Language (XML) 1.0. <http://www.w3.org/TR/REC-xml> (2000)
19. W3C. XML-Schema Part 1: Structures 2nd Ed. <http://www.w3.org/TR/xmlschema-1> (2004)
20. OMG. Response to the UML 2.0 OCL. <http://www.omg.org/docs/ad/02-05-09.pdf> (2002)

## Le diagnostic de fautes multiples

Lotfi Bouziani, Pr. Hafid Haffaf

Département d'Informatique, Université d'Es-Senia, Oran  
lotfibouziani@yahoo.fr, haffaf\_hafid@yahoo.fr

**Abstract.** Le diagnostic de fautes est un domaine de recherche très vaste et le cas des multi-fautes est très intéressant vu les effets et les interactions que peuvent avoir les différentes fautes. L'approche FDI peut être étendue pour prendre en considération les fautes multiples, mais les solutions obtenues restent toutes d'ordre combinatoire. Comme contribution, une solution basée sur l'optimisation combinatoire est présentée pour la résolution en temps polynomial du problème.

**Mots clés:** Diagnostic, FDI, RRA, Matrice d'incidence, Optimisation combinatoire, Algorithme génétique.

### 1. Introduction

Le diagnostic est un thème de recherche fédérant différentes communautés scientifiques (Automatique, Informatique, Productique...)

Le diagnostic a pour but d'établir un lien entre un symptôme observé, la défaillance ou la panne qui est survenue et ses causes.

Au sein de la communauté « Automatique », le diagnostic se retrouve sous l'appellation « Fault Detection and Isolation » ou (FDI), regroupant à la fois la détection d'une déviation de comportement qui donne lieu à la génération d'un symptôme (fonction détection) et l'isolation de la défaillance qui mène à la localisation de l'élément responsable de cette défaillance (fonction diagnostic). Cette tâche peut être décomposée en trois étapes (Fig. 1) :

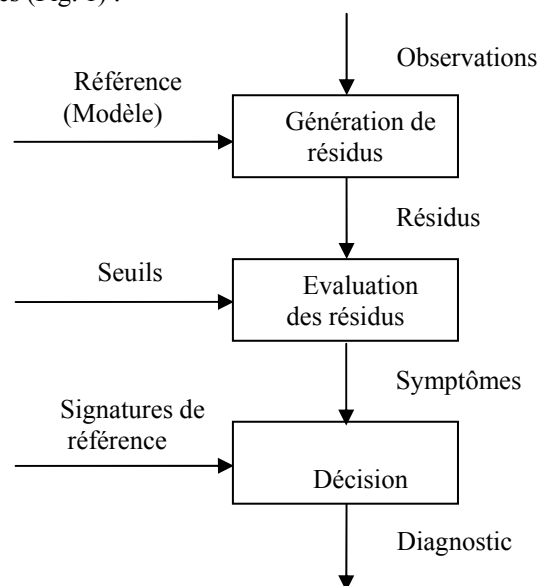


Fig. 1. Structure générale d'un système de diagnostic utilisant l'approche FDI [6].

1. La génération de résidus : qui permettent d'évaluer une différence par rapport aux conditions normales de fonctionnement ;
2. L'évaluation de ces résidus : les résidus sont comparés avec des limites définies préalablement et de cette comparaison résulte des symptômes générant un vecteur de cohérence. Un seul résidu suffit

pour connaître l'existence d'un défaut au sein du système, plusieurs résidus sont nécessaires afin de trouver la cause du défaut;

3. La décision : Consiste à comparer le vecteur de cohérence avec l'ensemble des signatures de référence, de bon et de mauvais fonctionnement rassemblées dans un tableau de signatures, appelé également matrice de diagnostic.

## 2. Le diagnostic par redondance analytique

Le diagnostic par redondance analytique suit une démarche qui peut se résumer comme suit :

1. La Détection : on dispose d'un modèle (mathématique) décrivant le fonctionnement normal du système et on surveille le fonctionnement réel en testant la cohérence entre ce modèle et les observations disponibles. Si celles-ci ne vérifient pas les équations du modèle, on en déduit que le fonctionnement réel n'est pas normal. On produit alors un symptôme (appelé aussi résidu).
2. L'Isolation : on compare le résultat de génération des résidus aux vecteurs binaires des différents défauts ou fautes dont on possède le modèle de mauvais fonctionnement. L'isolation du défaut se réalise lorsque la signature des symptômes correspond à l'une des signatures de panne.

Pour arriver à cette phase d'isolation (ou localisation), le diagnostic par redondance analytique a besoin de construire des tests de détection qui reposent sur des relations de redondance analytique (RRA). Ces derniers sont des modèles ne faisant intervenir que des variables mesurables ou connues du système. Ces modèles doivent permettre de calculer des résidus.

Une RRA peut être utilisée pour tester les composants physiques qui constituent son support (c.-à-d. les variables qu'elle utilise). Le but des RRA est de permettre de construire des tests de cohérence afin de générer des résidus qui permettront de décider si le fonctionnement s'apparente ou non à un fonctionnement normal.

Un autre concept défini dans l'approche FDI est celui des signatures des défauts. La signature théorique d'un défaut peut être envisagée comme la trace attendue du défaut sur les différents RRA qui modélisent le système [3]. Les valeurs des résidus sont comparées à des seuils et chaque décision issue de ces tests conduit à une valeur booléenne (0 : la valeur du résidu est en dessous du seuil ; 1 : la valeur du résidu a dépassé le seuil fixé). L'ensemble de ces valeurs booléennes forme un vecteur binaire appelé signature de faute. Ce vecteur binaire est comparé aux différentes signatures de panne ainsi qu'à la signature de fonctionnement normal. Cette comparaison conduit alors à une conclusion sur l'état du système : fonctionnement normal, défaillance identifiée ou finalement défaillance non identifiée.

La localisation qui suit la détection peut être effectuée à l'aide de la table de signature (matrice d'incidence). Les colonnes de cette table sont représentatives des différents défauts et les lignes des différents résidus. Les ensembles de diagnostics dans l'approche FDI sont donnés en termes de défauts présents dans la table de signature. La génération des ensembles de diagnostic est basée sur une interprétation des colonnes de la table de signature et consiste à comparer la signature des observations avec celle des défauts.

Pour que tous les défauts puissent être détectés, aucune colonne de la matrice des signatures théoriques de défauts ne doit être nulle, et pour que tous les défauts puissent être localisés, toutes les signatures théoriques doivent être distinctes.

On distingue donc, trois types de matrices d'incidence (Fig. 2) :

- non localisante : une colonne est nulle ou au moins deux colonnes sont identiques ;
- faiblement localisante : les colonnes sont non nulles et distinctes deux à deux ;
- fortement localisante : en plus d'être faiblement localisante, aucune colonne ne peut être obtenue à partir d'une autre en remplaçant un « 1 » par un « 0 » et inversement.

	$f_1$	$f_2$	$f_3$
$r_1$	1	1	1
$r_2$	1	1	1
$r_3$	1	0	0

Non localisante

	$f_1$	$f_2$	$f_3$
$r_1$	1	1	1
$r_2$	1	0	1
$r_3$	1	1	0

Faiblement localisante

	$f_1$	$f_2$	$f_3$
$r_1$	1	1	0
$r_2$	1	0	1
$r_3$	0	1	1

Fortement localisante

**Fig. 2.** Trois types de matrices d'incidences [3].

### 3. Le diagnostic multi-fautes

Dans le cas du diagnostic de fautes multiples, la matrice de diagnostic doit inclure les signatures des fautes multiples afin que ceux là soient considérés lors de la phase de décision.

Cependant, le nombre de possibilités évolue exponentiellement avec le nombre de fautes, d'une part et d'autre part, les fautes multiples sont plus difficiles à détecter, parce qu'elles peuvent masquer ou compenser leurs effets mutuellement [2].

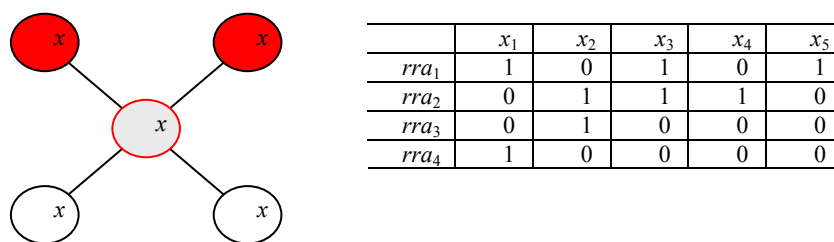
Les algorithmes de diagnostic de fautes multiples sont donc plus complexes que ceux des fautes simples pour deux raisons :

- d'abord, les effets d'une faute pourraient être masqués ou compensés (on dit aussi protégés) par les effets d'une autre faute.
- ensuite, les mêmes fautes multiples peuvent se manifester de différentes manières selon l'ordre dans lequel elles se sont produites.

Différentes approches de diagnostic multi-fautes existent, particulièrement dans la littérature IA. Une approche « classique » basée sur les modèles, considère le diagnostic comme une inférence logique, ou, en général, un problème de satisfaction de contraintes, où le diagnostic consiste à trouver une nomination satisfaisante aux composants du système (l'état normal ou l'état de panne), étant donné un jeu de contraintes décrivant le système et un jeu d'observations [4]. Cette approche est basée sur l'identification de conflits et la génération de candidats et le système proposé est appelé GDE pour General Diagnostic Engine.

Considérons quelques exemples qui démontrent les limitations des approches classiques dans des scénarios multi-fautes [5]. D'abord, une combinaison de fautes multiples peut produire un symptôme similaire à un autre symptôme de faute simple. Par exemple, considérons le réseau dans la figure Fig. 3 et supposons que les deux nœuds,  $x_4$  et  $x_5$  sont en panne. Alors, la signature de faute obtenue par sommation des signatures de  $x_4$  et  $x_5$  est  $(1, 1, 0, 0)^T$ . Cette signature est la même que celle de la panne de  $x_3$ , et il n'y a aucune façon de distinguer cette faute des fautes simultanées de  $x_4$  et  $x_5$ . Ainsi, la matrice d'incidence entraînera trois erreurs : l'oubli de deux fautes (2 erreurs de faux-négatives) et la localisation d'un composant « innocent »  $x_3$  (1 erreur de faux-positif).

Une autre situation de faute se produit quand une signature de fautes multiples est absente de la matrice d'incidence, comme par exemple, la colonne  $(1, 1, 0, 1)^T$  qui correspond à la panne de  $x_1$  et  $x_4$ . La matrice retournera  $x_3$  comme faute, puisque sa signature est la plus proche de  $(1, 1, 0, 1)^T$  selon la distance de Hamming. La distance de Hamming entre deux composants représente le nombre de bits différents entre leurs signatures respectives.



**Fig. 3.** Un système de réseau simple et sa matrice d'incidence ; plusieurs exemples de situations de multi-faute peuvent causer des erreurs de diagnostic.

Un autre problème peut survenir par exemple, à cause de contraintes d'ordre topologique. Ainsi, quelques situations multi-fautes peuvent être non-reconnaissables parce que quelques composants peuvent devenir « protégés » par les pannes d'autres composants [5]. Par exemple, dans la Fig. 3, la panne du nœud  $x_3$  protège  $x_4$  et  $x_5$  et il est impossible de savoir si  $x_4$  et  $x_5$  sont en panne.

#### 4. Le traitement des fautes multiples

Sous l'hypothèse de linéarité, l'isolation de fautes multiples peut être traitée en employant une matrice d'incidence étendue, incluant une nouvelle colonne pour chaque combinaison possible de fautes, menant à une solution combinatoire [7].

Le nombre de colonnes de la matrice d'incidence étendue ainsi obtenue, peut être égal à  $2^n - 1$  si toutes les fautes multiples sont considérées ( $n$  étant le nombre de composants simples du système).

La signature théorique d'une faute multiple est généralement obtenue à partir des signatures des fautes simples. Considérons que  $F_j$  est une faute multiple correspondant à l'occurrence simultanée de  $k$  fautes simples  $F_1, \dots, F_k$ , alors, les entrées du vecteur de signature de  $F_j$  est donnée par [7] :

$$\begin{cases} s_{ij} = 0 \text{ si } s_{i1} = s_{i2} = \dots = s_{ik} = 0 \\ s_{ij} = 1 \text{ sinon, c'est à dire si } \exists l \in \{1, \dots, k\} \text{ tel que } s_{il} = 1 \end{cases} \quad (1)$$

De plus, le fait que les signatures des fautes multiples soient obtenues à partir des signatures de fautes simples, exprime l'idée intuitive qu'une faute multiple ne peut affecter une RRA que si et seulement si, au moins une des fautes simples dont elle se compose affecte cette RRA. Cela signifie que la portée d'une faute multiple est l'union des portées de ses constituants en fautes simples.

L'interprétation des signatures de fautes multiples est la même que celle des fautes simples. Ceci implique que l'occurrence simultanée de plusieurs fautes produisant des situations de compensation n'est pas observable. Ceci est appelé « la supposition de non compensation multi-fautes » et qui est justifiée par la faible probabilité d'occurrence de tels cas [1].

Pour une meilleure isolabilité du système, une autre possibilité serait de trouver (ou prévoir) autant de RRA's que nécessaire pour couvrir toutes les combinaisons de fautes multiples. En effet, la signature de chaque composants (simple ou multiple) doit être différente et donc, le nombre de RRA's nécessaires devrait être au moins égal au nombre de composants simples du système. Par exemple, pour un système composé de 4 éléments simples  $\{C1, C2, C3, C4\}$ , on obtient 15 (ou  $2^4 - 1$ ) éléments multiples  $\{C1, C2, C3, C4, C1C2, C1C3, C1C4, C2C3, C2C4, C3C4, C1C2C3, C1C2C4, C1C3C4, C2C3C4, C1C2C3C4\}$  et pour avoir une signature de faute différente pour chaque élément, on doit avoir au moins 4 RRA's, c'est-à-dire le même nombre d'éléments simples. L'obtention de nouvelles RRA's n'étant pas toujours possible et les signatures de la matrice d'incidence étendue étant souvent similaires, d'autres techniques d'optimisation sont utilisées afin d'améliorer l'isolabilité du système.

On peut, par exemple, regrouper les éléments simples ou multiples ayant les mêmes signatures en une seule colonne. Ceci permettra de réduire considérablement le nombre de colonnes de la matrice d'incidence et ainsi de réduire le temps nécessaire pour la parcourir. Dans l'exemple de la figure Fig. 4, une matrice de 10 composants est étendue aux 1032 colonnes possibles, puis elle est réduite à seulement 15 colonnes différentes.

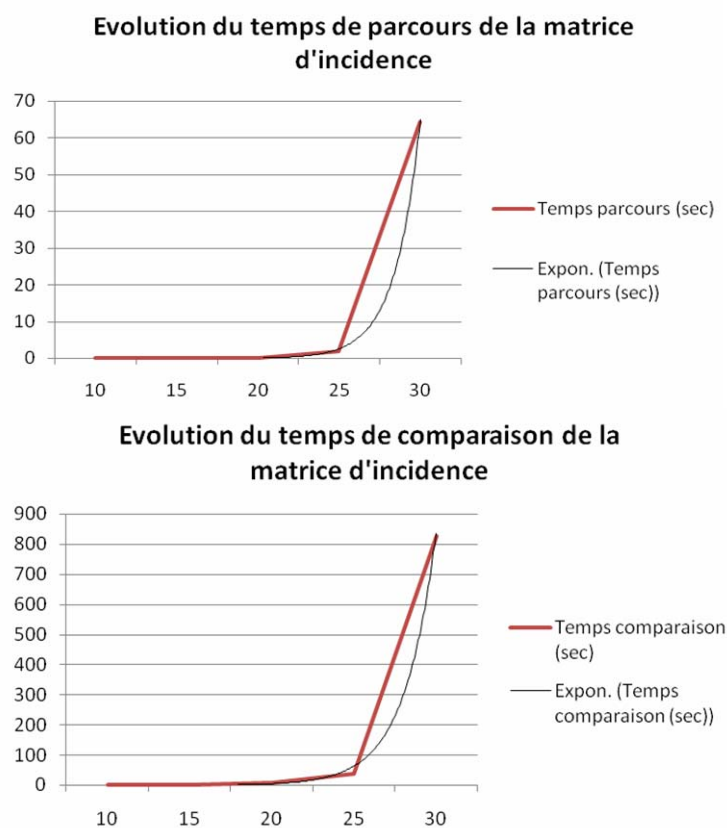
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
RAA1	1	0	0	0	1	0	0	1	0	1
RAA2	0	1	0	0	1	1	0	0	1	0
RAA3	0	0	1	0	0	1	1	1	0	0
RAA4	0	0	0	1	0	0	1	0	1	1

	C1	C2	C3	C4	{C5, C1- C2, C1- C5, C2- C1- C2- C5}	{C6, C2- C3, C2- C6, C3- C2- C3- C6}	{C7, C3- C4, C3- C7, C4- C3- C4- C7}	{C8, C1- C3, C1- C8, C3- C1- C3- C8}	{C9, C2- C4, C2- C9, C4- C2- C4- C9}	{C10, C1- C4, C1- C10, C4- C1- C4- C10}	{C1- C6, C7, C1- C6, C7- C8- C10}	{C1- C7, C1- C7, C3- C4- C5- C8- C10}	{C1- C9, C1- C2- C4- C5- C9- C10}	{C2- C7, C2- C3- C4- C6- C7- C9}	{C5- C7, C8- C9, C1- C2- C3- C4- C5- C6}
RR41	1	0	0	0	1	0	0	1	0	1	1	1	1	0	1
RR42	0	1	0	0	1	1	0	0	1	0	1	0	1	1	1
RR43	0	0	1	0	0	1	1	1	0	0	1	1	0	1	1
RR44	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1

**Fig. 4.** Après optimisation et à partir de 1023 signatures, seules 15 sont réellement différentes.

## 5. Application

Toutes ces techniques ont été implémentées sur un logiciel afin de mieux comprendre les enjeux et les notions traitées dans cet article. Après plusieurs essais nous avons obtenus les résultats illustrés par la figure Fig. 5. Ces résultats montrent bien l'aspect combinatoire de la méthode de diagnostic FDI lorsque les combinaisons de fautes multiples sont considérées.



**Fig. 5.** Evolution du temps de parcours de la matrice d'incidence ainsi que le temps nécessaire pour la comparer avec la signature de faute observée.

## 6. Solution proposée

A partir de ces constatations, l'idée nous est venue d'utiliser les techniques de l'optimisation combinatoire pour générer et travailler sur cette matrice et trouver les signatures qui correspondent le plus à la signature de faute rencontrée.

Parmi ces techniques, nous prévoyons d'utiliser les algorithmes génétiques. Un algorithme génétique utilise des techniques évolutives observées en biologie pour trouver des solutions aux problèmes complexes de recherche et d'optimisation. Un algorithme génétique va faire évoluer une population constituée de plusieurs solutions dans le but d'en améliorer les individus et c'est donc, à chaque génération, un ensemble d'individus qui sera mis en avant et non un individu particulier. Nous obtiendrons donc un ensemble de solutions pour un problème et pas une solution unique. Les solutions trouvées seront généralement différentes, mais seront d'une qualité équivalente.

Dans notre cas, nous allons essayer de trouver les solutions qui se rapprochent le plus de la signature de faute observée, et pour cela nous allons commencer par une population de fautes simples. Leurs signatures seront comparées avec l'observation et celles qui s'en approchent le plus (selon la distance de Hamming) seront sélectionnées.

Ensuite, un croisement entre fautes sera effectué afin de générer de nouvelles signatures qui seront à leur tour comparées et les meilleures seront sélectionnées. Cette opération doit être répétée un nombre de fois défini dans le paramétrage de l'algorithme. D'autres points sont à paramétrer tels que le choix et la taille de la solution initiale.

Remarquons que le but de l'algorithme n'est pas de trouver toutes les solutions possibles, comme dans l'approche FDI étendue, mais de trouver des solutions qui se rapprochent le plus de la signature de l'observation et qui font intervenir le moins de composants possible. En effet, plus il y a de composants qui rentrent dans la génération d'une signature, plus l'effet de protection est amplifié et donc plus on s'éloigne de la solution réelle. Ces deux objectifs définissent la fonction objective de l'algorithme génétique qu'on souhaite utiliser.

Prenons comme exemple la matrice d'incidence d'un système à 10 composants simples et la signature de faute observée de la figure Fig. 6.

	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C4</i>	<i>C5</i>	<i>C6</i>	<i>C7</i>	<i>C8</i>	<i>C9</i>	<i>C10</i>	<i>F</i>
<i>RA1</i>	1	0	0	0	1	0	0	1	0	1	1
<i>RA2</i>	0	1	0	0	1	1	0	0	1	0	1
<i>RA3</i>	0	0	1	0	0	1	1	1	0	0	0
<i>RA4</i>	0	0	0	1	0	0	1	0	1	1	1

**Fig. 6.** La matrice d'incidence d'un système à 10 composants simples et la signature de faute observée.

La solution proposée peut se résumer ainsi (voir le tableau Tab. 1) :

1. Parcourir la matrice à la recherche des signatures qui se rapprochent le plus de la signature de faute observée. Les solutions trouvées vont être utilisées comme population initiale pour l'algorithme génétique. Dans notre exemple, les signatures trouvées sont celles des composants {*C1*, *C5*, *C10*}.
2. Enrichir les signatures initiales par d'autres signatures afin de se rapprocher encore plus de la signature de faute observée. Pour cela, d'autres composants sont ajoutés et leurs signatures sont fusionnées avec les premières. Pour notre exemple, les nouvelles signatures sont celles de : {*C1C2*, *C5C4*, *C10C6*}.
3. Cette étape est répétée et en ajoutant et en retirant des composants afin que les signatures obtenues se rapprochent le plus de la solution. Le fait de trouver des solutions exactes ne signifie pas l'arrêt d'exécution, car nous sommes devant une situation où plusieurs solutions totalement différentes peuvent exister.
4. La condition d'arrêt est un certain nombre d'itérations défini comme paramètre. Dans notre exemple, nous nous sommes arrêtés à 4 itérations (en réalité, un algorithme génétique doit être répété au moins 100 fois).

**Tab. 1.** Exemple des résultats obtenus avec un algorithme génétique.

Itérations	Sol. 1	Sol. 2	Sol. 3
Initial.	C1	C5	C10
1 <sup>ère</sup>	C1C2	<b>C5C4</b>	C10C6
2 <sup>ème</sup>	C1C2C3	C4C3	C10C7
3 <sup>ème</sup>	<b>C1C2C4</b>	C4C2	C10C7C8
4 <sup>ème</sup>	C1C2C5	C4C2C3	<b>C10C7C9</b>

## 7. Conclusion

Nous avons présenté dans cet article l'approche de diagnostic de fautes appelée FDI et nous nous sommes intéressé particulièrement au cas des fautes multiples. Nous avons ensuite montré que la méthode FDI dans le cas de fautes multiples converge vers un problème d'ordre combinatoire. Enfin nous avons proposé d'utiliser une approche qui utilise les algorithmes d'optimisation combinatoire, tels que les algorithmes génétiques.

## References

1. Cordier, M-O., Dague, P., Lévy, F., Montmain, J., Staroswiecki, M., Travé-Massuyès, L.: Analyse Comparative des Approches de l'IA et de l'Automatique Appliquées au Diagnostic à Base de Modèles. Journée S3/IMALAIA (2002).
2. Daigle, M., Koutsoukos, X., Biswas, G.: Multiple Fault Diagnosis in Complex Physical Systems. Vanderbilt University. Nashville, TN 37235 (2006).
3. Gertler, J.: Analytical Redundancy Methods in Fault Detection and Isolation: Survey and Synthesis. IFAC Symposium on Fault Detection Supervision and Safety (1991).
4. de Kleer, J., Williams, B.C.: Diagnosing Multiple Faults. Xerox Palo Alto Research Center, USA (2006).
5. Odintsova, N., Rish, I., Ma, S.: Multi-fault Diagnosis in Dynamic Systems. IBM T.J. Watson Research Center (2003).
6. Theilliol, D.: Contribution à l'Etude et au Développement des Systèmes Tolérants aux Défauts : Diagnostic et Accommodation à Base de Modèles Linéaires et au-delà. Thèse de Doctorat, Université Nancy 1, France (2003).
7. Weber, P., Gentil, S., Ripoll, P., Foulloy, L.: Multiple Fault Detection and Isolation. 14th World Congress of IFAC. Beijing, China (1999).



*Mobile & Wireless technology*

---



## **Enhancing DSR Protocol in MANETS**

Sofiane Boukli Hacene<sup>1</sup>, Ahmed Lehireche<sup>1</sup>, Ahmed Meddahi<sup>2</sup>

<sup>1</sup>Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS,  
Computer Science Department, Sidi Bel Abbes University, Algeria

<sup>2</sup>Département informatique et réseaux, ENIC Telecom  
Université LILLE 1, France

<sup>1</sup>boukli@univ-sba.dz, <sup>1</sup>elhir@yahoo.com, <sup>2</sup>meddahi@enic.fr

**Abstract.** A mobile ad-hoc network (MANET) is a collection of nodes, which are able to communicate forming an arbitrary and dynamic network using a wireless medium. To reach any node in the network a multihop path is necessary. Due to node mobility and other factors many links become unusable at any time, therefore, network topology change dynamically. Several routing protocols exist allowing and facilitating communication between mobile nodes. Routing protocols must be able to respond rapidly to topological changes and use only valid routing information. One of the promising routing protocols is DSR (Dynamic Source Routing). This protocol presents some problems. The major problem in DSR is that the route cache contains some inconsistent routing information; this is due to node mobility. This problem generates longer delays for data packets. In order to reduce the delays we propose a technique based on cleaning route caches for nodes within an active route. Our approach has been implemented and tested in the well known network simulator GLOMOSIM and the simulation results shows that protocol performance have been enhanced.

**Keywords:** Mobile Ad hoc networks, DSR, Routing protocols, simulation and performance evaluation.

### **1 Introduction**

A MANET is a collection of nodes, which are able to connect on a wireless medium forming an arbitrary and dynamic network. Each node in a MANET is a router. Communication between nodes requires a multihop wireless path from a source to a destination, so nodes must cooperate in routing operation. All nodes are mobile and can be connected dynamically in an arbitrary manner to form a network [1]. A challenge in the design of ad hoc networks is development of dynamic routing protocols that can efficiently find routes between two nodes. The key task of routing protocols is to deliver packet from the source node to the given destination [2]. The existing routing protocols are, traditionally, divided into two classes, depending on when a node acquires a route to a destination. Reactive protocols invoke a route discovery procedure on demand only. Thus, when a route is needed, some sort of flooding-based global search procedure is employed. One of the promising reactive routing protocols is DSR. In general, routing protocol presents some problems, and one of the major problems in DSR is longer data packets delays caused by the search process in the cache. In this paper we propose a technique to solve this problem.

The remainder of the paper is organized as follows. First we give an overview on DSR and its operations. This is followed by focusing on stale routes problem in section 3. In section 4, a presentation of different works that tries to solve this problem is given. Next, we concentrate on the proposed technique to solve the problem of delay caused by stale routes in cache in section 5. In section 6, we present the performance evolution of the proposed approaches, and finally we conclude in section 7.

### **2 Dynamic Source Routing Protocol (DSR)**

Dynamic Source Routing [3] is based on source routing, where the source node specifies the whole path to destination node in the packet header. When a source node needs to communicate with a destination node, it first searches in its route cache for a route to the destination, if a route is found, the source node uses it, otherwise the source node initiate route discovery mechanism to discover a route. In a route discovery mechanism the source node floods a route request message (RREQ) to neighboring nodes. The

message contains the source address, the destination address, the request id, and a list containing the complete path to destination. When node receives this request, it proceeds as following:

- a) If the node has seen this same request before, it ignored the request.
- b) If the receiving node is the destination itself or a node having a route to the destination in its cache, it returns a route reply message (RREP), which contains: the source address, the destination address, and the route record in the route request message. The route reply message is sent back to the source node by following the same route record in route request message in reverse order.
- d) Otherwise, it appends its own address to the route record, and rebroadcast the route request message to its neighboring nodes.

When the source node receives the route reply message, it starts sending data packets to the destination. When a route failure happens, the node upstream the broken link sends back to the effective source a route error message (RERR). Nodes receiving RERR message remove broken link from its routes cache. The source node initiate a route discovery if it receives RERR message, it still need a route to the destination and no alternate route in its cache.

### **3 The Stale Routes Problem in the DSR Protocol**

The DSR has the advantage of learning routes by scanning for information in packets that are received. A route from A to C through B means that A learns the route to C, but also that it will learn the route to B. The source route will also mean that B learns the route to A and C and that C learns the route to A and B. This form of active learning is very good and reduces overhead in the network, by this way each node in DSR can find alternative routes when link failure happens. This propriety will have a bad repercussion on route cache when node mobility is high. The route caches will contain in this case many stale routes to destinations that may be used to reach a destination and this generates longer delay for data packets. Several previous studies deal with stale routes problem [4, 5].

### **4 Related Work**

In this section we will present some ideas to enhance the DSR routing protocol.

Chen and Hou in [4] used a neighbor link-state information exchange mechanism. Once a connection has been established, the neighbor link-state information is exchanged among nodes along the route from the source to the destination. As the information of the neighbor lists is piggybacked in data packets, the nodes on the source route are able to learn the partial topology around the neighborhood of the connection. The simulation results show that with limited overhead incurred in neighbor list dissemination, the proposed protocol outperforms DSR with either path or link caches in term of packet delivery ratio and route discovery overhead.

In [6], He et al propose an active packet technique to improve DSR. The mean idea is allowing a packet to visit each node twice, this packet is named "Active packet". The objective of the first visit is to obtain topology information of the network; and the objective of the second visit is to update route caches according to the obtained information. In the header active packet header contains a marker field to indicate if the packet is in the first or the second visit. The payload of the active packet is a connection matrix for the network topology. The active packet is generated periodically. Simulation results show that the method reduced the miss rates by up to 60% and routing packet numbers by up to 47%.

An enhancement to DSR by using a link breakage prediction algorithm was proposed in [7]. A mobile node uses signal power strength from the received packets to predict the link breakage time, and sends a warning to the source node only if the link is soon-to-be-broken. The source node can perform a pro-active route rebuild to avoid disconnection. Simulation results show that the method reduced the dropped packets (by at least 20%). The tradeoff is an increase in the number of control messages by at most 33.5%.

## 5 Proposed technique

In order to the delay experienced by data packets and reduce stale routes in caches, we add an expiration time for each route inserted in the cache. This idea is inspired from route management in the routing table of AODV [8,9]. When learning new routes, a node must set an expiration time for each route insert in the cache, and when this time expires, the route is removed from the route cache of the node. Each time a route is used, its expiration time is set. The max value is fixed to 10 Seconds (represent approximately 1% of simulation time) empirically.

## 6 Performance Evaluation & Simulation Results

In order to evaluate the effectiveness of the proposed technique described above, we add it to the basic version 3 of DSR available in the GLOMOSIM simulator and we compare it with the original version using performance metrics.

The simulation environment and the performance metrics used will be described in the next paragraph, the simulation results presentation and discussion is done later.

### 6.1 Simulation Environment

We have used the implementation of DSR version 3 included in the well known GlomoSim simulator. Our results are based on the simulation of 50 wireless nodes forming an ad hoc network moving about in a rectangular area of 1500 meters by 300 meters for 900 seconds of simulated time. The source-destination pairs are spread randomly over the simulation area, sending 4 data packets per second following a CBR (constant bit rate) fashion. For our simulation 10-20-30 and 40 source-destination pairs are chosen. Traffic sessions are established randomly and stay active until the simulation ends. A random waypoint mobility model [3] is used. The movement scenario we used for each simulation is characterized by a pause time. Each node begins the simulation by selecting a random destination in the simulation area and moving to that destination at a speed distributed uniformly between 0 and 20 meters per second. It then remains stationary for pause time seconds. This scenario is repeated for the duration of the simulation. We carry out simulations with movement patterns generated for 10 different pause times starting by 0s varying by 100s Step until 900s (the length of the simulation) is reached, which corresponds to limited motion.

The physical radio characteristics of each mobile node's network interface, such as the antenna gain, transmission power, and receiver sensitivity, were chosen to approximate the Lucent WaveLAN direct sequence spread spectrum radio[10].

The performance metrics [11] used to evaluate performance are:

**Average end-to-end delay of data packets:** This includes all possible delays caused by buffering during route discovery, queuing at the interface queue, retransmission delays at the MAC layer, and propagation and transfer times.

**Communication overhead** is the total number of control packets, and including route request, route reply, and route error packets generated for each delivered data packet.

**Number of broken link** is the number of invalid routes for sending data across it; the proposed technique reduces the use the broken links.

### 6.2 Simulation Results and Discussions

We report the results of the simulation experiments for the original DSR protocol and for Optimized DSR (DSR Opt). In all figures below, Pause time varied between 0 seconds and 900 seconds. When pause time is 0 seconds this denotes high mobility, while 900 seconds pause time means no mobility. Each scenario is repeated 5 times and the average values of the results are chosen.

### Broken Links

It is observed that the number of broken links is high when node mobility is high (Fig. 1, 2, 3 and 4). The number of sources also affects the number of broken links. When the number of sources increases the number of broken links also increases because the need of more routes to destinations. It can be noticed from those figures that DSR Opt results in substantially fewer link breaks, especially when pause times are small. This is due to the expiration time mechanism added to DSR; consequently the probability of using a stale route is minimized (The protocol tends to use fresher routes).

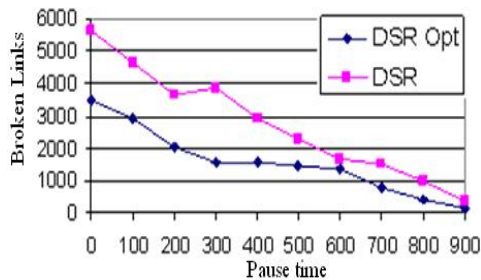


Fig. 1. Number of Broken Links for 10 sources

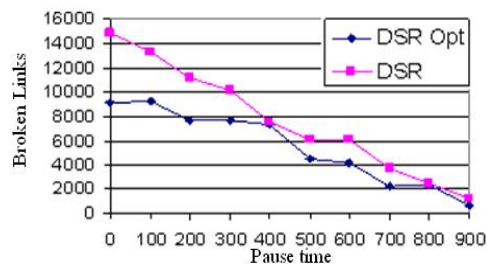


Fig. 2. Number of Broken Links for 20 sources

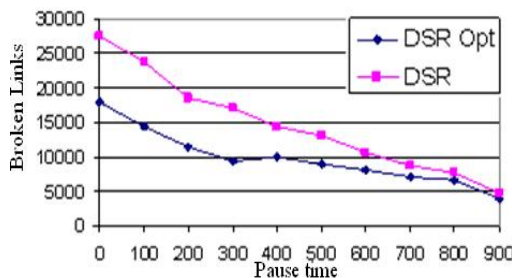


Fig. 3. Number of Broken Links for 30 sources

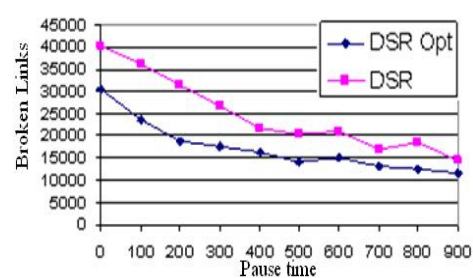


Fig. 4. Number of Broken Links for 40 sources

### Average End-to-End delay

In Fig.5, 6, 7 and 8 the results obtained for the end-to-end delay metric are presented. We observe that the end-to-end delay increases significantly when the number of sources increases. Minimizing broken links contribute directly to minimizing end to end delay for data packets. When a broken link happens data packets experience a delay in DSR Opt lower then in DSR Because of the reduced number of cached route. The results show that DSR Opt outperforms DSR significantly when the number of sources is low and node motion is high. This enhancement of DSR is suitable for Multimedia flows which can not tolerate higher delays.

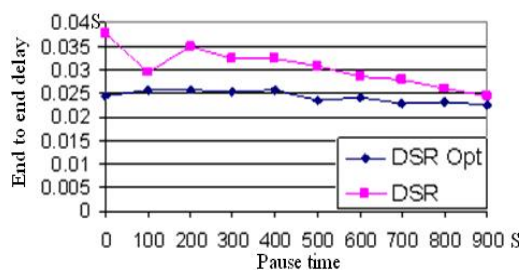


Fig. 5. Average End-to-End delay for 10 sources

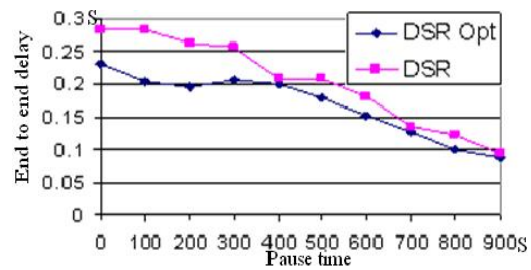


Fig. 6. Average End-to-End delay for 20 sources

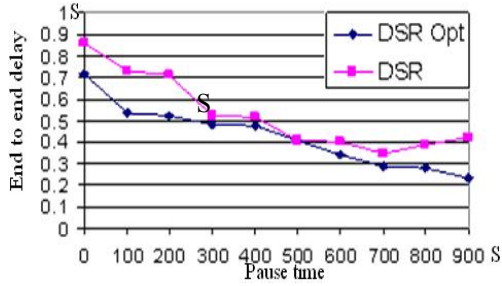


Fig. 7. Average End-to-End delay for 30 sources

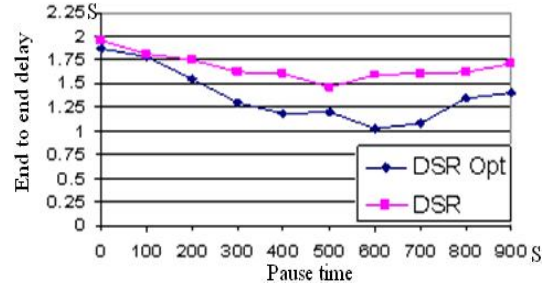


Fig. 8. Average End-to-End delay for 40 sources

### Communication Overhead

Fig. 9, 10, 11 and 12 show how mobility and number of sources affect the communication overhead. We notice that communication overhead is high when node motion is high; this is due to the dynamic topology of the network. It is also observed that the overhead is high when the number of sources is high. This results from the fact that many sources try to discover routes to destinations, which increase the number of control packets and so the communication overhead. The results show that DSR Opt results in substantially less overhead when the mobility is moderate (100 and 400s of pause time); this has a good impact on energy because the number of control packets generated is low. Sometimes, DSR opt generate higher overhead than DSR, this can be explain by the fact that when using the expiration time technique, some valid routes may be removed from cache, which generate a new route discovery.

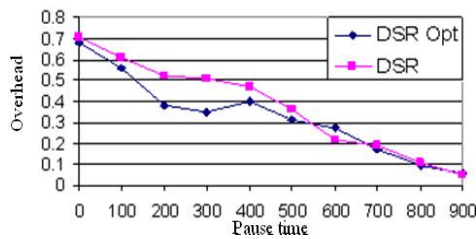


Fig. 9. Communication Overhead for 10 sources

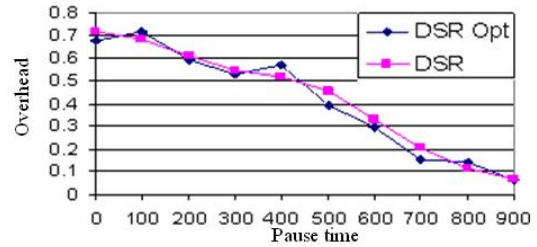


Fig. 10. Communication Overhead for 20 sources

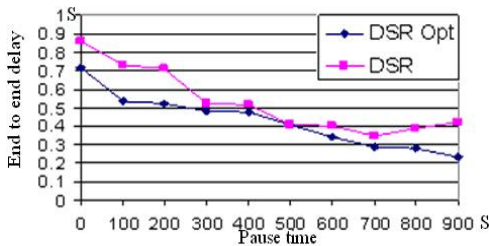


Fig. 11. Communication Overhead for 30 sources

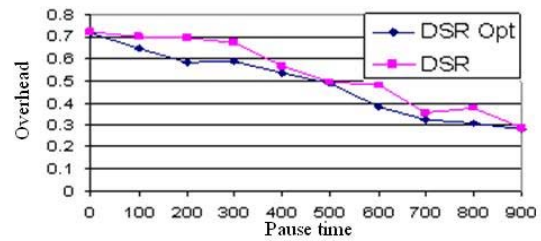


Fig. 12. Communication Overhead for 40 sources

## 7 Conclusion

In this paper, we have improved the promising DSR routing protocol for ad hoc networks. We have equipped DSR with expiration time technique for routes in route cache. This technique has been inspired from route management in the routing table of AODV routing protocol, in order to avoid the use of stale route in routing.

The performance of the proposed technique was evaluated and compared with DSR using detailed simulations. Several common performance metrics were considered. The simulation results show that the proposed algorithm performs well; it can overall generate lower communication overhead, fewer broken links and lower Average end-to-end delay.

## References

1. Amitava, M. Somprakash, B. Debashis, S.: Location Management and Routing in Mobile Wireless Networks. Wireless internet handbook: technologies, standards, and application, CRC Press, Inc (2003) 351-353
2. Ramjee, R. Marina M.: Technology Trends in Wireless Communications, Artech House Inc (2003)
3. Johnson, D. B. Maltz, D.: A Dynamic Source Routing in Ad Hoc Wireless Networks. Draft-ietf-manet-dsr-09, INTERNET DRAFT (1996)
4. Chen, W. and Hou, J. C.: Dynamic, Ad-Hoc Source Routing with Connection-Aware Like-State Exchange and Differentiation. In: Proc. IEEE Globecom, Vol. 1. IEEE Globecom NEW YORK (2002) 188-194
5. Mathur, A.: Performance Analysis and Enhancement on Dynamic Source Routing for Mobile Ad Hoc Networks. PHD Thesis in Computers Science, University of Texas at San Antonio (2005)
6. He, Y. Raghavendra, C. S. Berson, S. Braden, B.: Active Packets Improve Dynamic Source Routing for Ad-Hoc Networks. Next Generation Teletraffic and Wired/Wireless Advanced Networking. Lecture Notes in Computer Science, Vol. 4712. Springer-Verlag, Berlin Heidelberg New York (2007) 367–378
7. Qin, L. Kunz, T.: Increasing Packet Delivery Ratio in DSR by link Prediction. In: Proc. 36th Annual Hawaii International Conference on System Sciences (HICSS'03), Vol. 9 IEEE Computer Society (2003) 300.1
8. Belding-Royer, E. Perkins, C.: Evolution and future directions of the ad hoc on demand distance-vector routing protocol. Ad Hoc Networks Journal, Vol.1, Elsevier B.V, (2003) 125–150
9. Belding-Royer, E. Perkins, C. Das, S.: Ad hoc On-Demand Distance Vector (AODV) Routing. draft-ietf-manet-aodv-13.txt, INTERNET DRAFT (2003)
10. Tuch, B.: Development of WaveLAN, an ISM Band Wireless LAN. AT&T Technical Journal, Vol.4, (1993) 27–33
11. Corson, S. Macker, J.: Mobile Ad Hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations. IETF, (1999)

## Amélioration des Interactions entre les Protocoles Transport et MAC dans les MANET.

Hamrioui Sofiane<sup>1</sup>, Lalam Mustapha<sup>1</sup>

<sup>1</sup> LARI, Université Mouloud Mammeri de Tizi Ouzou, Algérie.  
s.hamrioui@gmail.com

**Résumé.** L'informatique mobile est entrée dans une phase de maturité, grâce à une floraison de produits permettant d'accéder à distance à Internet ou à des données personnelles ou d'entreprise. Pour ne pas rester en marge de cette réalité et affronter les nouveaux défis auxquels sont confrontés les réseaux de 3<sup>ème</sup> et 4<sup>ème</sup> générations, on s'investit dans leur étude en particulier ceux de type MANET (Mobile Ad hoc Network) afin de contribuer à l'introduction de certains mécanismes assurant une qualité de services (QoS) à ces derniers. Parmi les paramètres de QoS que l'on va étudier, on cite la performance du protocole TCP. Nous présenterons dans ce papier une amélioration des interactions entre le protocole MAC et celui de transport (notamment TCP) dans les MANET en proposant un nouvel algorithme de Backoff du protocole MAC et un nouveau mécanisme d'acquittement du protocole TCP.

**Mots clés:** MANET, Protocole, Transport, MAC, QoS.

### 1 Introduction

La performance du protocole TCP [1] est un paramètre de QoS [2] [3] très important et qui fait, de plus en plus, l'objet de plusieurs évaluations. Cette dernière se dégrade dans les MANET [4] vu que le protocole TCP est confronté à de nouveaux défis [5]. Plusieurs solutions [6] ont été proposées afin d'essayer d'y remédier bien mais elles sont peu adaptées au protocole MAC (Medium Acces Control) [7] employé et limitées à un protocole sur une couche sans aucune interaction avec les autres couches. Ces interactions entre le protocole MAC et celui du transport (notamment TCP) devront être étudiées afin d'adapter le comportement de ces derniers pour assurer la bonne QoS supporté par les MANET. Notre contribution consiste à étudier, en premier lieu, les conséquences de la non prise en compte des interactions entre le protocole de la couche transport (TCP) et le protocole MAC. A travers des simulations réalisées sous l'environnement NS (Network Simulator) [8], nous avons mis en lumière le problème de la dégradation de la performance de TCP. Après avoir exposé la problématique, nous avons proposé deux solutions pour lesquelles on a étudié leurs effets sur la performance du MANET.

### 2 Interactions des protocoles MAC et TCP

#### 2.1 Analyse du problème

Plusieurs solutions ont été proposées dans la littérature pour améliorer la performance de TCP dans le contexte des MANET [6]. Ces solutions sont peu adaptées au protocole MAC employé et sont dans la plupart limitées à un protocole sur une couche sans aucune interaction avec les autres couches. Dans cette section, nous illustrons l'impact de la couche MAC sur la performance du protocole TCP.

**Impact du noeud exposé.** Dans la "Fig. 1", on considère une topologie en chaîne formée de  $n$  nœuds dans laquelle seulement une seule connexion TCP est établie. On suppose que c'est le nœud 1 qui commence de transmettre pour le nœud  $n$ .



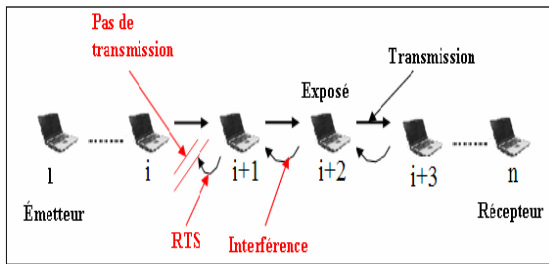


Fig. 1. Impact du noeud exposé sur TCP

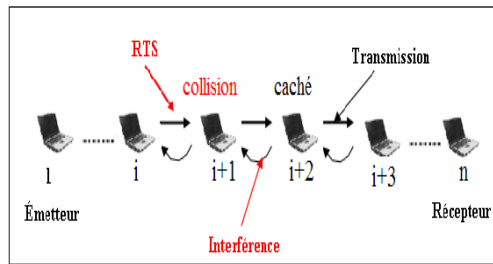


Fig. 2. Impact du noeud caché sur TCP

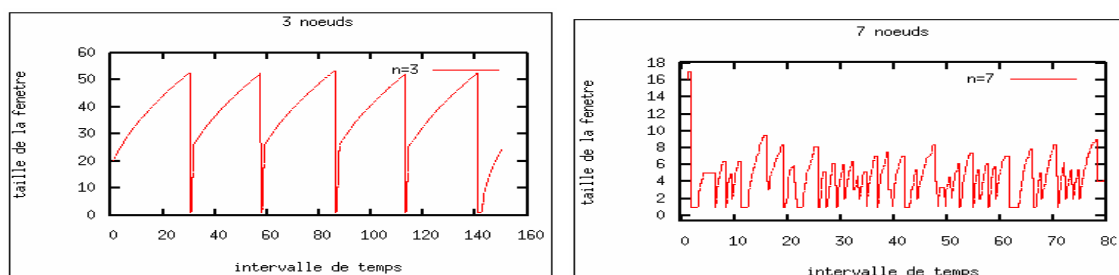
Supposons que le nœud  $i+1$  possède une trame de donnée (qui est dans ce cas un acquittement TCP) à envoyer pour le nœud  $i$ . Étant donné que le nœud  $i+2$ , situé dans la portée du nœud  $i+1$ , est en cours de transmission pour le nœud  $i+3$ , il l'empêche de transmettre sa trame. Le nœud  $i+1$  doit alors attendre la fin de la transmission du nœud  $i+2$  pour le nœud  $i+3$  et rentre ainsi en contention pour le médium. Ici le nœud  $i+2$  est le nœud exposé (par rapport au nœud  $i+1$ ). Cette attente pour que le médium se libère peut durer très longtemps surtout si le nœud  $i+2$  transmet continuellement des trames de données (c.a.d des segments TCP), cela inhibe le nœud  $i+1$  d'accéder au médium. Cette impossibilité pour le nœud  $i+1$  de transmettre la trame vers le nœud  $i$  revient à dire que, au niveau TCP, c'est un acquittement TCP d'un paquet émis avec succès vers la destination qui sera empêché d'arriver au nœud destinataire. Mais après une certaine durée (timeout), le nœud  $i$  (l'initiateur de la connexion TCP) considère que le paquet envoyé est perdu et associe cela à une congestion, cela résulte le déclenchement du mécanisme d'évitement de congestion, qui aura comme conséquence la diminution du débit TCP (fenêtre TCP).

**Impact du nœud caché.** Traitons maintenant ("Fig. 2") le problème du nœud caché. On pose même ici les mêmes hypothèses de la figure "Fig. 1".

Supposons que le nœud  $i$  désire émettre une trame de donnée pour le nœud  $i+i$ , alors d'après le principe de MAC 802.11 DCF, le nœud  $i$  tente d'envoyer un RTS puis attend une réponse par le nœud  $i+3$  avec un CTS. Mais le nœud  $i+1$  ne reçoit jamais le RTS parce que ce paquet rentre en collision avec l'interférence causée par le nœud  $i+2$  considéré ici comme nœud caché. Après plusieurs tentatives d'accès au médium échouées, l'algorithme de backoff au niveau MAC augmente la taille de la fenêtre de contention, ainsi les chances pour le nœud  $i$  d'avoir le médium sont minimales. Donc, le nœud  $i$  va croire qu'il s'agit d'une rupture de route alors il informe le nœud source et ce dernier initie de ce fait une procédure de découverte de nouvelle route. Avant que la nouvelle route soit établie, aucun paquet de donnée ne pourra être envoyé. Durant ce processus, l'émetteur TCP doit attendre et ainsi évoquer le contrôle de congestion s'il constate que le délai (timeout) a expiré ce qui conduira également à une diminution de débit.

## 2.2 Simulation de scénarios sous NS

Dans nos simulations, le modèle de propagation utilisé est le modèle two-ray ground, pour le protocole MAC on a opté pour le 802.11 IEEE et un modèle d'antenne omnidirectionnelle de l'outil NS. Comme protocole de routage nous avons utilisé AODV, une taille de file d'interface de 50 dans chaque nœud. Nous avons testé la version NewReno de TCP. Nous avons simulé quatre scénarios : 3, 7, 9 et 12 nœuds. La taille des paquets de données TCP est de 1040 bytes. Dans les scripts, l'option *macTrace* est activée dans le but d'avoir les détails sur les trames du protocole MAC, cela nous a permis d'analyser la raison de chaque perte de paquet TCP qui s'est produite. Les résultats de simulations des scénarios avec  $n = 3, 7, 9$  et 12 sont donnés ci-dessus.


Fig. 3. L'évolution de la taille de la fenêtre TCP standard pour  $n=3$  nœuds et  $n=7$  nœuds.

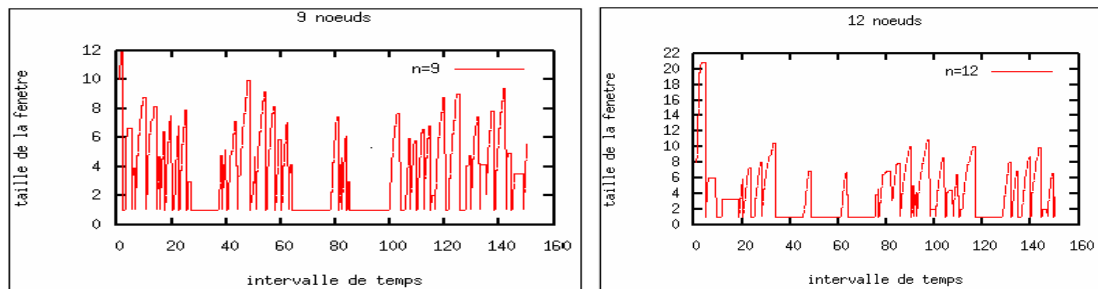


Fig. 4. L'évolution de la taille de la fenêtre TCP standard pour  $n=9$  nœuds et  $n=12$  nœuds.

Nous remarquons que dans le cas où  $n=3$  nœuds le graphe reflète exactement le fonctionnement du protocole TCP dans des conditions idéales. Nous distinguons facilement une même sous courbe répétée plusieurs fois. A son tour, cette courbe est divisée en deux étapes séparées par un seuil égal à 25 sur le graphe. Ces deux étapes correspondent respectivement au *départ lent* et *évitement de congestion* dont le passage d'une étape à l'autre dépend seulement de la congestion du médium. En analysant le fichier trace obtenu après simulations dans lequel il y a absence presque totale de pertes de RTS/CTS. Pour les scénarios  $n=7$  jusqu'à  $n=12$  nœuds, nous constatons que plus le nombre de nœuds participants dans la chaîne augmente, plus les pertes de paquets TCP sont fréquentes et ainsi plus la taille de la fenêtre TCP actuelle diminue. On analysant les fichiers trace obtenus, nous avons constaté que ces pertes de paquets TCP sont dues aux pertes des trames de contrôle (RTS/CTS) utilisé par le protocole MAC 802.11 pour initier des échanges de trames de données, ce qui conduira à une perte des paquets correspondants. Après expiration du timeout, TCP procédera à la régularisation de son débit qui réduira la taille de sa fenêtre TCP.

### 2.3 Solutions proposées

**1) Retarder les acquittements TCP et favoriser les paquets de données TCP.** La fiabilité fournie par TCP consiste à s'assurer que chaque paquet émis par l'émetteur sera bien transmis à l'application destinataire grâce à un système d'acquittements. Ces acquittements TCP peuvent rentrer en collision avec leurs paquets de données correspondants. C'est sur ce principe que repose cette première solution, elle consiste à essayer de retarder les paquets d'acquittements TCP dans le but de réduire les collisions causées par ces derniers et leurs paquets TCP. Nous donnons ci-après les résultats de simulations des scénarios avec 9 et 12 nœuds.

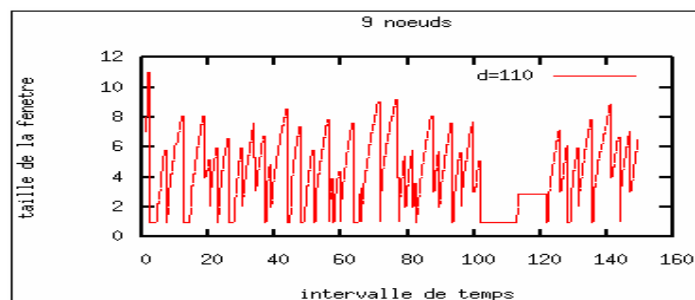


Fig. 5. L'évolution de la taille de la fenêtre TCP avec  $d=110$  dans le cas  $n=9$  nœuds.

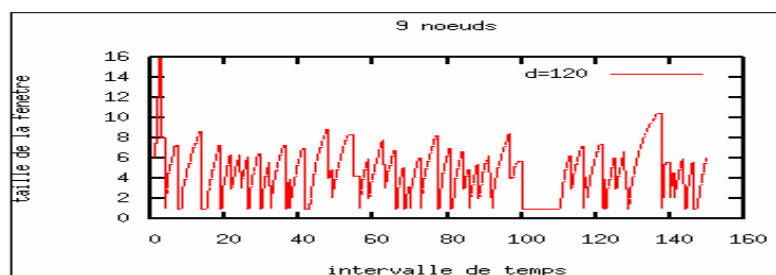


Fig. 6. L'évolution de la taille de la fenêtre de TCP avec  $d=120$  ms dans le cas  $n=9$  nœuds.

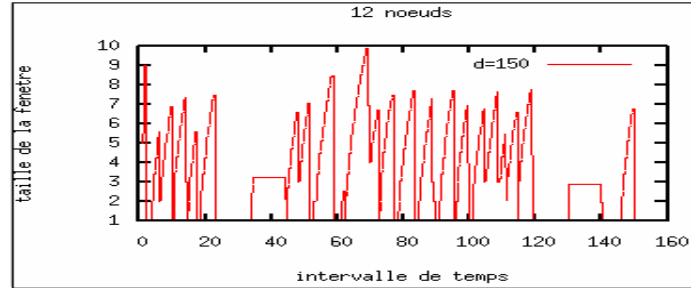


Fig.7. L'évolution de la taille de la fenêtre de TCP avec  $d=150$  ms dans le cas  $n=12$  nœuds.

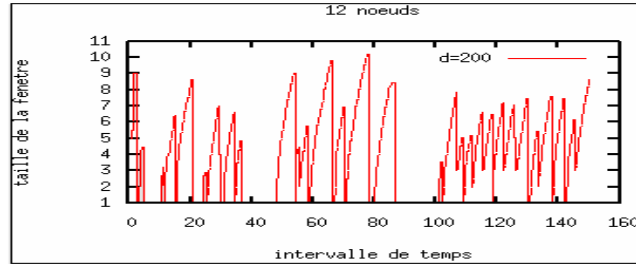


Fig. 8. L'évolution de la taille de fenêtre de TCP avec  $d=200$  ms dans le cas  $n=12$  nœuds.

Nous remarquons une nette amélioration de la performance de TCP des figures "Fig. 5" et Fig. 6" par rapport à la figure "Fig. 4" ( $n=9$  nœuds) et des figures "Fig. 7" et "Fig. 8" par rapport à la figure "Fig. 4" ( $n=12$  nœuds) sur le plan perte de paquets qui est réduite et la taille de la fenêtre qui est augmentée relativement. Nous constatons également que utiliser des délais raisonnablement grands ( $d=120$ ms pour le cas  $n=9$  nœuds et  $d=200$ ms pour le cas  $n=12$  nœuds) pour le retard d'acquittements améliore mieux TCP par rapport aux petites valeurs de ces derniers. Cette amélioration est due au fait que dans cette version de TCP les collisions causées par les paquets de données TCP avec leurs acquittements correspondants sont réduites.

**Modèle analytique de la solution :** Après avoir réalisé plusieurs scénarios avec des nombres différents de nœuds, nous avons relevé pour un échantillon de scénarios (entre 6 et 32 nœuds) les valeurs optimales des retards d'acquittements, ces dernières nous ont permis de dresser le graphe de la figure "Fig. 9" suivant :

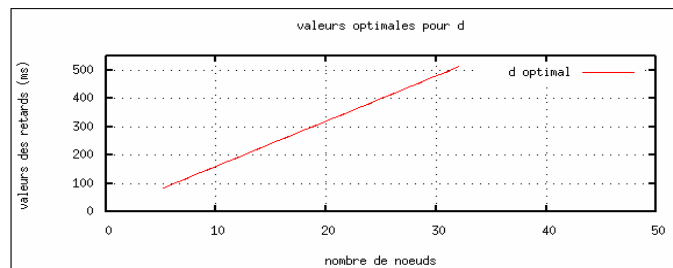


Fig. 9. Les valeurs optimales de retard d'acquittements pour les scénarios entre 6 et 32 nœuds.

Le graphe correspond à une droite dont la fonction est de la forme :

$$d = y(n) = \alpha n \quad (1)$$

tel que  $n$  est le nombre de nœuds utilisés dans le scénario.

Pour déterminer la valeur de  $\alpha$ , on utilise la tangente de l'angle de la droite, on aura donc :

$$\text{Tg}(\text{angle}) = dx/dy = 15.62 \quad (2)$$

La formule (1) devient :

$$d = y(n) = 15.62 n \quad (3)$$

Pour la topologie choisit, nous avons que cette formule n'est performante que pour des valeurs de délais inférieur ou égales à 500. Nous fixons comme objectif de nos proches travaux de réutiliser cette formule avec des topologies différentes et de déterminer le lien entre la constante  $\alpha$  et la topologie choisit ainsi que entre cette dernière et les facteurs de l'environnement de communication.

**2) Adapter l'algorithme de Backoff du protocole MAC.** Le protocole MAC repose sur l'algorithme de backoff qui lui permet de déterminer qui va accéder au médium sans fil tout en évitant les collisions. Cet algorithme prend en compte une fenêtre de temps appelé *fenêtre de contention (Contention Windows)*. Sa valeur change entre une  $CW_{min}$  minimale et une  $CW_{max}$  maximale. Elle augmente avec le nombre de collisions à partir de  $CW_{min}$  jusqu'à  $CW_{max}$ . Nous proposons dans cette deuxième suggestion d'adapter l'algorithme de backoff au nombre de nœuds intervenants dans la communication. Le principe est d'essayer de déterminer des valeurs pivots (optimales) pour  $CW_{min}$  et  $CW_{max}$  qui réduiront les collisions entre nœuds. Nous avons gardé les mêmes paramètres que précédemment, sachant que le protocole TCP utilisé est toujours la version Newreno. Les résultats obtenus avec 9 nœuds et 12 nœuds sont donnés ci-dessus.

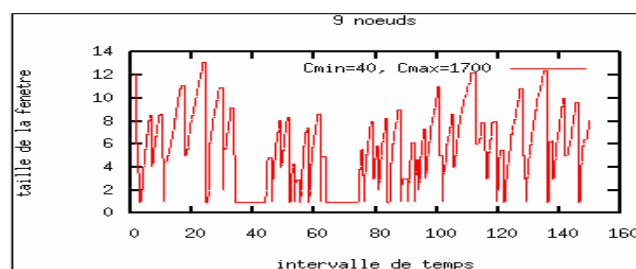


Fig. 10. L'évolution de la taille de la fenêtre TCP avec  $CW_{min}=40$  et  $CW_{max}=1700$  dans le cas  $n=9$  nœuds.

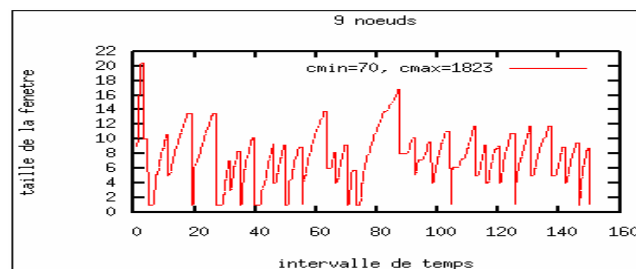


Fig. 11. L'évolution de la taille de la fenêtre TCP avec  $CW_{min}=70$  et  $CW_{max}=1823$  dans le cas  $n=9$  nœuds.

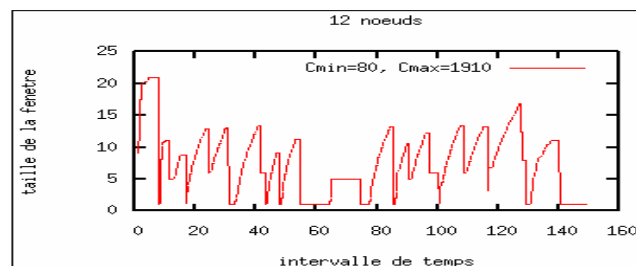


Fig. 12. L'évolution de la taille de la fenêtre TCP avec  $CW_{min}=80$  et  $CW_{max}=1910$  dans le cas  $n=12$  nœuds.

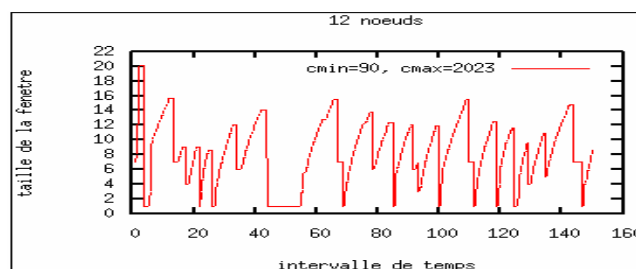
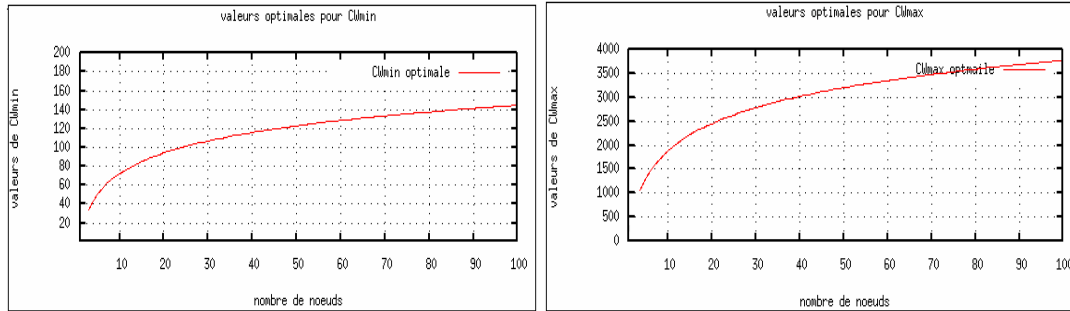


Fig. 13. L'évolution de la taille de la fenêtre TCP avec  $CW_{min}=90$  et  $CW_{max}=2023$  dans le cas  $n=12$  nœuds.

Les figures “Fig. 10” et “Fig. 11” montrent une nette amélioration de la performance de TCP par rapport à la figure “Fig. 4” ( $n=9$  nœuds) car les pertes de paquets sont réduites et la taille actuelle de la fenêtre est augmentée. Nous rappelons que les valeurs de  $CWmin$  et de  $CWmax$  utilisées dans le scénario associé à la figure “Fig. 4” sont celles fixées par défaut à savoir  $CWmin=31$  et  $CWmax=1023$ . Cette amélioration revient au fait qu’on a augmenté les valeurs des bornes de la fenêtre de contention des nœuds ce qui réduira la probabilité d’avoir les mêmes chance d’accéder au médium sans fil ainsi réduire les collisions qui conduira ainsi à diminuer leurs conséquences sur le protocole TCP ainsi améliorer sa performance. Les figures “Fig. 12” et “Fig. 13” montrent aussi une amélioration de la performance de TCP par rapport à la figure “Fig. 4” ( $n=12$  nœuds). Après avoir fixé  $CWmin$  et  $CWmax$  à des valeurs différentes durant plusieurs scénarios, nous avons constaté que les valeurs 70 pour  $CWmin$  et 1823 pour  $CWmax$  donne le meilleur résultat pour le scénario à 9 nœuds. Quant au scénario à 12 nœuds, les valeurs dégagées comme optimales sont  $CWmin=90$  et  $CWmax=2023$ . Ces valeurs sont relativement plus grandes que celle utilisées dans le scénario avec 9 nœuds à cause de l’augmentation du nombre de nœuds qui est de 12 nœuds. Donc il faut toujours trouver un bon compromis entre les valeurs de ces dernières et le nombre de nœuds utilisé.

**Modèle analytique de la solution :** Après avoir réalisé plusieurs scénarios avec des nombres différents de nœuds communiquant, nous avons relevé pour chaque scénario les valeurs optimales  $CWmin$  et  $CWmax$ , ces dernières nous ont permis de dresser les deux graphes illustrés sur la figure “Fig. 14” qui correspondent à la relation qui existent les valeurs optimales de  $CWmin$  et  $CWmax$  et celle du nombre du nœuds.



**Fig. 14.** Les valeurs optimales de  $CWmin$  et de  $CWmax$  pour les scénarios entre 3 et 100 nœuds.

Les graphes de la figure “Fig. 14” correspondent clairement à des courbes d’une fonction logarithmique translaté d’une constante, alors on pose les fonctions de la forme suivante:

$$CWmin = y(n) = \beta \log(n) \quad (4)$$

Tel que  $n$  est le nombre de nœuds employés dans le scénario.

Pour déterminer la valeur de la constante  $\beta$ , nous procédons comme suite :

$$\beta = y(n)/\log(n) = 31. \quad (5)$$

31 correspond bien à la valeur initiale de  $CWmin$ .

$$CWmax = y(n) = \gamma \log(n) \quad (6)$$

Tel que  $n$  est le nombre de nœuds employés dans le scénario.

déterminons la valeur de la constante  $\gamma$  comme suite :

$$\gamma = y(n)/\log(n) = 820 = 1024 * 5/4 \quad (7)$$

1024 correspond bien à la valeur initiale de  $CW_{max}$ . Le nouveau Backoff est alors :

$$n \leftarrow n + 1$$

$$CW(n) = (aCW_{min}(m) + 1) * 2^n - 1$$

**Tel que :**

$$CW_{min} \leq CW(n) \leq CW_{max} \quad (8)$$

$$CW_{min}(m) = \beta \log(m)$$

$$CW_{max}(m) = \gamma \log(m)$$

**Avec :**

$m$  est le nombre de nœuds employés

$\beta$  est une constante égale à  $31 = CW_{min}$  initiale

$\gamma$  est une constante égale à  $1024 * 5/4 = CW_{max}$  initiale  $* 5/4$

Nous fixons comme objectif de nos proches travaux de réutiliser cet algorithme avec des topologies différentes et de déterminer le lien entre les constantes  $\beta$  et  $\gamma$  et la topologie choisit ainsi que entre ces dernières et les facteurs de l'environnement de communication.

### 3 Conclusion et perspectives

Dans le cadre de ce papier, nous avons proposé deux optimisations de l'algorithme de Backoff du protocole MAC et celui d'acquiescement du protocole TCP pour améliorer les interactions qui existe entre eux. Nous sommes passé ensuite à l'étude des effets de cette amélioration sur la performance du MANET. Les résultats obtenus étaient satisfaisant et montrent une nette amélioration de la performance du protocole TCP en particulier et du MANET en général. Notre travail ne constitue qu'une partie de ce qui doit se faire dans le contexte de l'amélioration de la performance d'un MANET en intervenant sur les interactions entres protocoles. Nous fixons comme objectif de nos proches travaux de réutiliser cet algorithme avec des topologies différentes et de déterminer le lien entre les constantes  $\alpha$ ,  $\beta$  et  $\gamma$  et la topologie choisit ainsi que entre ces dernières et les facteurs de l'environnement de communication.

### References

1. M.W. Murhammer, O. Atakan, S. Bretz, L.R. Pugh, K. Suzuki, and D.H. Wood: TCP/IP Tutorial and Technical Overview, International Technical Support Organization, 1998.
2. J. Li.: Quality of Service (QoS) Provisioning in Multihop Ad Hoc Networks, Doctorate of Philosophy, Computer Science in the Office of Graduate Studies, University of California, 2006.
3. O. Fouial. : Découverte et fourniture de services adaptatifs dans les environnements mobiles, Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications Paris, 2004.
4. S. Basagni, M. Conti, S. Giordano and I. Stojmenovic. : Mobile Ad hoc Networking, IEEE Press, Wiley-Interscience A John Wiley & Sons, Inc., Publication, 2004.
5. A.A. Hanbali, E. Altman, and P. Nain.: A Survey of TCP over Ad Hoc Networks, INRIA, June 2005.
6. R.D. Oliveira, and T. Braun.: TCP in Wireless Mobile Ad Hoc Networks, NCCR-MICS project, 2002.
7. P. Mohapatra, and S.V. Kkishnamurthy.: Ad hoc Networks Technologies and Protocols, Springer Science + Business Media, Inc, 2005.
8. <http://www.isi.edu/nsnam/ns/ns-documentation>.

# Une Nouvelle Approche pour l'Economie d'Energie dans les Réseaux Sans Fil Mobiles

Hamrioui Sofiane

Département d'Informatique, Université Mouloud Mammeri de Tizi Ouzou, Algérie  
hamriouirayane@yahoo.fr

**Résumé.** La conservation d'énergie est importante dans un réseau sans fil mobile pour augmenter la durée de vie des noeuds et, par conséquent, celle de tout le réseau. Cette tâche est particulièrement difficile du moment que les actions de la conservation d'énergie doivent être effectuées de manière distribuée. Cet objectif est d'autant plus difficile à atteindre à cause, d'une part, de la miniaturisation des composants qui conduit à une miniaturisation des batteries et, d'autre part, des nouvelles générations de terminaux mobiles qui intègrent des composants sophistiqués offrant des performances plus importantes mais, en contrepartie, consommant plus d'énergie. Nous présenterons dans ce présent papier une nouvelle approche pour l'économie de l'énergie au niveau de la couche réseau. Cette nouvelle approche hybride se veut de reprendre et de synthétiser les atouts des approches existantes ainsi que d'éviter leurs inconvénients. Elle se base essentiellement sur une qualité énergétique d'un lien entre deux terminaux mobiles, et est ainsi adapté pour un protocole à état des liens.

**Mots clés:** Réseaux sans fil mobiles, Economie d'Energie, OLSR.

## 1 Introduction

Le domaine des réseaux sans fil mobiles suscite, de plus en plus, d'intérêt ces dernières années. La particularité de ce type réseau est qu'ils sont faciles et rapides à déployer et ils peuvent opérer de façon autonome ou être connectés à d'autres types de réseaux. L'une des majeures problématiques dans les réseaux sans fil mobiles reste au niveau de l'économie d'énergie. Le but de notre travail est d'établir une nouvelle approche pour l'économie de l'énergie au niveau de la couche réseau et de l'implémenter sur un protocole de routage. Cette nouvelle approche hybride se base essentiellement sur une qualité énergétique d'un lien entre deux terminaux mobiles.

## 2 Approche proposée pour l'économie d'énergie

Plusieurs solutions ont été développées pour économiser l'énergie d'un terminal mobile et ainsi augmenter sa durée de vie. Certaines de ces approches sont utilisées de nos jours, tandis que d'autre reste au niveau expérimentale. Une classification des différents approches développés sont très bien détaillées dans [1] [2].

### 2.1 Préliminaires

Représentons un réseau sans fil mobile par un graphe  $G = (V, E)$  ou :

$V$  est l'ensemble des nœuds.

$E \in V^2$  l'ensemble des arrêtes donnant les communications possibles entre les nœuds.

a) **Le voisinage d'un nœud** [3]

Soit  $N(u)$  l'ensemble du voisinage d'un nœud  $u$  :

$$N(u) = \{v \in V / v \neq u \text{ et } (u, v) \in E\} \quad (2)$$

b) **Ajustement de portée** [3] [4]

L'utilisation de l'interface radio engendre un coût énorme pour la consommation du mobile.

Fenney propose plusieurs analyse [5] [6] montrant le surcoût énergétique des interfaces radio et l'accroissement important de l'énergie consommée en fonction de la puissance ou portée de l'émission (les deux termes sont fortement liée).

Soit  $R$  la portée maximal d'un nœud et  $u \in V$ . L'ajustement de portée consiste à assigner un facteur  $f_u$  de manière à utiliser une portée de communication.

$$r(u) = f_u \times R \quad (2)$$

Avec  $f_u$  valeur continue sur  $[0,1]$ .

#### c) Mesure de distance [4]

Pour pouvoir évaluer les distances entre les noeuds. On peut utiliser plusieurs techniques dont :

➤ **GPS** (ou *Global Positionning System*) : Système permettant de donner sa position par rapport au repère terrestre. Plusieurs satellites synchronisés envoient régulièrement des messages donnant leur position.

➤ **Mesure de puissance** : À la réception d'un signal radio, une interface réseau peut mesurer la puissance reçue. Si le noeud est capable d'associer à cette valeur une mesure de la distance, il peut alors déduire à quelle portée se situe l'émetteur du signal.

➤ **Différence de phase** : Une interface radio possédant deux antennes peut découvrir selon quel angle d'arrivée le signal radio d'un voisin lui est arrivé. Il lui suffit pour cela de calculer la différence de temps entre l'arrivée du signal sur chaque antenne.

#### d) Modélisation [3] [4]

On incluant l'ajustement de portée et la distance euclidienne séparant deux nœuds, on obtient une autre représentation d'un réseau de capteurs :

Soit  $G=(V,E)$  un réseau sans fil mobile et  $d(u,v)$  la distance euclidienne séparant  $u$  et  $v$ , avec  $u, v \in V$

Soit  $r(u)$  la portée de transmission d'un nœud  $u$  avec  $r(u) \leq R$

Le graphe induit par la fonction d'assignation de portée est défini par  $G_r=(V,E_r)$  tel que :

$$E_r = \{(u, v) \in V^2 \mid d(u,v) \leq r(u)\} \quad (3)$$

L'ensemble des voisins d'un nœud  $u$  devient alors :

$$N_r(u) = \{v \in V \mid v \neq u \text{ et } (u, v) \in E_r\} \quad (4)$$

#### e) Métriques et Lois de composition sur les métriques [7]

Soit  $G=(V,E)$  un graphe avec  $|V| = n$  noeuds et  $|E| = m$  arcs.

Soit **poid** <sub>$i,j$</sub>  une métrique (un poids) caractérisant le lien  $(i,j)$ .

La valeur du poids de tous chemin orienté  $C = (i, j, k, \dots, q, r)$  peut être une des compositions suivantes :

##### ➤ Métrique Additive :

On dit qu'une métrique **poid** est additive si :

$$\text{poid}(C) = \text{poid}_{i,j} + \text{poid}_{j,k} + \dots + \text{poid}_{q,r} \quad (5)$$

Les métriques délai, nombre de saut et le coût énergétique sont des métriques additives.

##### ➤ Métriques Multiplicative:

On dit qu'une métrique **poid** est multiplicative si :

$$\text{poid}(C) = \text{poid}_{i,j} * \text{poid}_{j,k} * \dots * \text{poid}_{q,r} \quad (6)$$

## 2.2 Description de la solution

La solution consiste à affecter des poids aux liens, et de les inclure lors du calcul du plus court chemin. Ces poids sont calculés à partir de l'énergie résiduelle et de la distance séparant chaque paire de nœuds. Voici les principales étapes à suivre pour calculer le poids d'un lien :

**Etape 1** : Définir un seuil de portée d'émission  $S$  en fonction de l'énergie résiduelle. Ceci permettra de partitionner le champ de communication en deux zones :

➤ **Zone proche** : La portée d'émission est comprise entre 0 et  $S$  et les communications dans cette zone sont prioritaires.



➤ **Zone éloignée** : La portée est comprise entre **S** et **R** ; **R** étant la portée d'émission maximal. Les communications dans cette zone sont moins prioritaires.

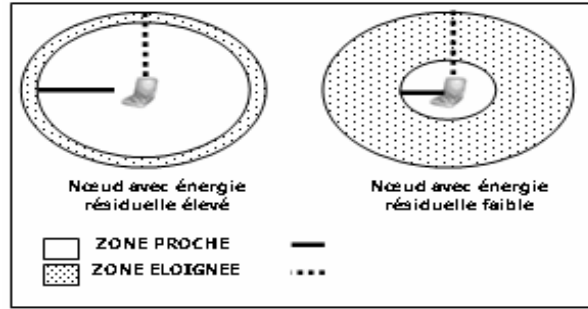


Fig. 1. Zone proche / zone

**Etape 2** : Classer les voisins en :

- **Voisins proches** : appartenant à la zone proche.
- **Voisins éloignés** : appartenant à la zone éloignée.

Ceci en leur attribuant un poids  $P_z$ .

Le poids  $P_{zi}$  du voisin  $v_i$  est calculé avec la fonction  $f_{pz}^S$

$$P_{zi} = f_{pz}^S(d(u, v_i)) \quad (7)$$

Les deux type de voisins sont illustrés sur la figure si dessous :

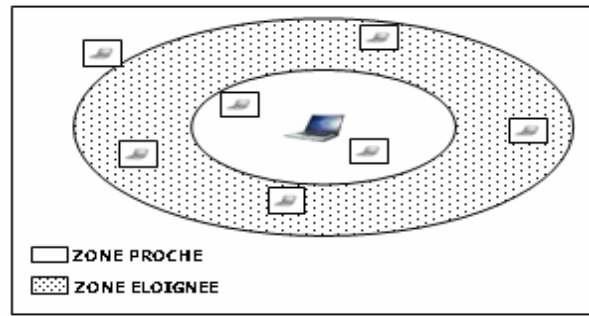


Fig. 2. Voisins proches / voisins

**Etape 3** : Attribuer un poids énergétique  $P_e$  à chaque voisin. Ce poids sera défini par le niveau de l'énergie résiduelle du voisin. Plus le niveau d'énergie est grand et plus le poids  $P_e$  est fort.

Le poids  $P_{ei}$  d'un nœud  $n_i$  est calculé avec la fonction  $f_{pe}$

$$P_{ei} = f_{pe}(\text{énergie résiduelle de } n_i) \quad (8)$$

**Etape 4** : Attribuer un poids  $P_i$  à chaque lien (arc)  $l_i$  liant l'émetteur à chaque voisin  $v_i$ . Le poids  $P_i$  est calculé avec la fonction  $f_{pi}$  tel que :

$$P_i = f_{pi}(P_z, P_e) \quad (9)$$

**Etape 5** : Inclure les poids  $P_i$  lors du calcul du plus court chemin. Ceci en utilisant la loi additive.

## 2.3 Implémentation de la solution

Nous avons implémenté notre nouvelle approche sur le protocole OLSR [8], nous présenterons dans ce qui suit l'environnement de la simulation en mettant l'axant sur les dispositifs que nous avons utilisé.

Les nouvelles données utilisées sont :

energy\_ : c'est l'énergie résiduelle d'un nœud à un instant donnée.

seuil\_ : c'est le seuil de la portée d'émission qui est relatif à energy\_ et à la portée maximale du nœud R

$$\text{seuil\_} = (\text{energy\_} \times R) / \text{énergie maximale} \quad (10)$$

Energie maximale est la quantité maximale que peut contenir la batterie d'un nœud.

**distance\_** : c'est la distance séparant deux nœuds.

*Remarque* : dans la simulation, la distance séparant deux nœud n1 et n2 sera calculé en utilisant leur coordonnées X et Y avec la formule suivante :

$$\text{distance\_} = \sqrt{(X_{n1} - X_{n2})^2 + (Y_{n1} - Y_{n2})^2} \quad (11)$$

pz\_ : poids de la zone, détermine si un nœud est dans la zone proche ou la zone éloigné.

$$\text{pz\_} = \begin{cases} 1 & \text{Si } \text{seuil\_} \geq \text{distance\_} \text{ (proche)} \\ 3 & \text{Si } \text{seuil\_} \leq \text{distance\_} \text{ (éloigné)} \end{cases} \quad (12)$$

pe\_ : poids énergétique, relatif à l'énergie résiduelle d'un nœud.

$$\text{pe\_} = \begin{cases} 1 & \text{Si } ((\text{energy\_} \times 100) / \text{énergie maximale}) \geq 50 \\ 3 & \text{Si } 20 \leq ((\text{energy\_} \times 100) / \text{énergie maximale}) \leq 50 \\ 5 & \text{Si } ((\text{energy\_} \times 100) / \text{énergie maximale}) \leq 20 \end{cases} \quad (13)$$

poid\_ : poids du lien.

$$\text{poid\_} = \text{pz\_} + \text{pe\_} \quad (12)$$

Ainsi les valeurs possibles du poids du lien sont :

$\text{pz\_} \backslash \text{pe\_}$	1	3
1	2	4
3	4	6
5	6	8

## 2.4 Simulations et interprétations

Le but principal de nos expérimentations est d'évaluer les performances de notre nouvelle approche. Pour ce faire nous avons utilisé l'outil NS (Network Simulator) [9] devrons au préalable étudier l'impact de cette dernière sur les performances du protocole OLSR, cette première étude portera sur la fiabilité des deux version (OLSR, E\_OLSR) et nous permettra de définir un scénario de simulation pour la seconde étude. La seconde étude portera sur les performances de notre approche, nous y étudierons la consommation de l'énergie dans le réseau ainsi que la durée de vie de ce dernier.

### Etude de la fiabilité

- *Variation de la vitesse :*

On fixe: Temps de pause = 50 sec, la charge du réseau = 20 sources.

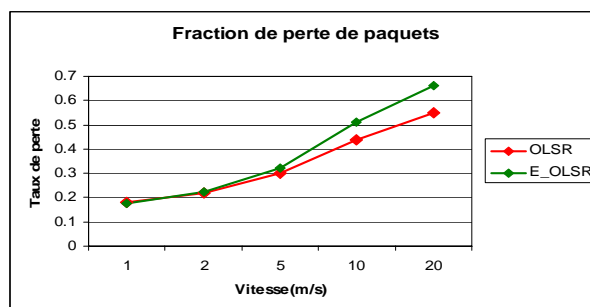


Fig. 3. Taux de perte de donnée en fonction de la vitesse

Pour de faibles vitesses des nœuds, le réseau présente une faible mobilité et une assez grande stabilité ; les défaillances des liens sont donc moins fréquentes que dans le cas d'une grande mobilité. En conséquence, la fraction de perte de données est plus petites pour de faibles vitesses de mouvement des nœuds, et croît avec l'augmentation de la mobilité des nœuds. Pour de faibles vitesses, le taux de perte de paquets des deux protocoles est presque identique, mais avec l'augmentation de la vitesse, le protocole OLSR offre de meilleurs résultats que E\_OLSR. Cela est du à la différence du nombre de sauts des chemins calculé par OLSR et E\_OLSR. En effet OLSR génère des chemins avec un nombre minimal de sauts. L'impact de la mobilité sur ces chemins est moins important que l'impact de la mobilité sur les chemins générés par E\_OLSR.

- **variation du temps de pause :**

On fixe : vitesse = 2m/s, charge du réseau = 20 sources

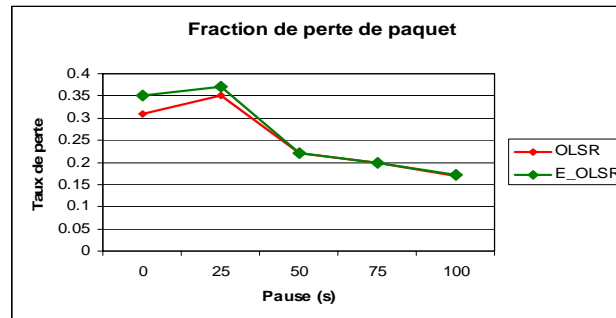


Fig. 4. Taux de perte de donnée en fonction du temps de pause

L'augmentation du temps de pause conduit à la stabilité du réseau, ce qui favorise l'arrivé des paquets de données à leur destination, donc la fraction de perte de données réduit. Pour des temps de pauses minimaux, OLSR présente de meilleurs résultats que E\_OLSR, puis avec l'augmentation du temps de pause les deux protocoles se rejoignent et offrent les mêmes résultats. Comme pour la variation de la vitesse, cette différence de performances lors de la variation du temps de pause est du à la longueur des chemins (en sauts). Nous concluons donc que l'implémentation de notre nouvelle approche n'est totalement réussie que pour une faible mobilité.

- **variation du nombre de sources :**

On fixe : Vitesse = 2m/s, temps de pause = 20 sources.

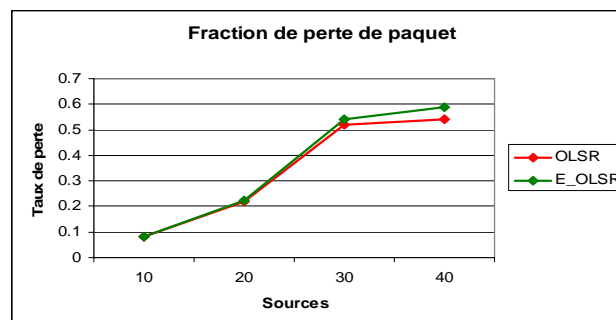


Fig. 5. Taux de perte de donnée en fonction du nombre de sources CBR

La figure montre que le taux de perte augmente relativement avec l'augmentation du nombre de sources CBR. En effet l'augmentation du nombre de connexion conduit à la surcharge du réseau, ce qui favorise les collisions des paquets et donc augmente le taux de perte de données.

Les chemins optimisés en nombre de sauts de OLSR réduisent le taux de perte de paquets par rapport aux chemins de E\_OLSR. Plus un paquet est retransmis, plus il risque une collision, et ainsi provoquer une perte de données.

Nous concluons donc que l'implémentation de notre nouvelle approche n'est totalement réussie que pour un nombre de source CBR inférieur à 50% du nombre des nœuds du réseau.

### Etude de la consommation de l'énergie

D'après les résultats de l'étude précédente, nous avons fixé le scénario suivant pour l'étude de la consommation de l'énergie.

Nombre de nœuds = 10 nœuds, nombre de sources CBR = 3 sources, vitesse maximal 2 m/s, temps pause 50 sec, durée simulation 1000 sec, énergie initiale 20 joule.

- **énergie totale du réseau**

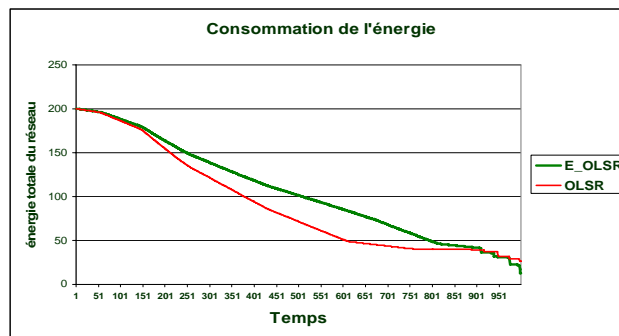


Fig. 6. L'évolution de l'énergie totale en fonction du temps

Sur ce graphique l'énergie totale du réseau OLSR décroît plus vite que celle du réseau E\_OLSR sur l'intervalle [0,600] sec. Puis sur l'intervalle [600,1000] sec, nous remarquons une stabilisation du niveau de l'énergie totale du réseau OLSR, cela est dû à la perte de connectivité du réseau. En effet à  $t = 600$  le nombre de nœud vivant du réseau OLSR est de 4, et ces 4 nœuds ne communiquent pas probablement à cause de l'éloignement. Sur le même intervalle, l'énergie du réseau E\_OLSR continue de décroître, preuve que les nœuds continuent de communiquer. Nous concluons à partir de l'évolution de l'énergie sur l'intervalle [0,600] que notre approche a réalisé une économie de 14% sur l'énergie totale du réseau.

- **Durée de vie du réseau**

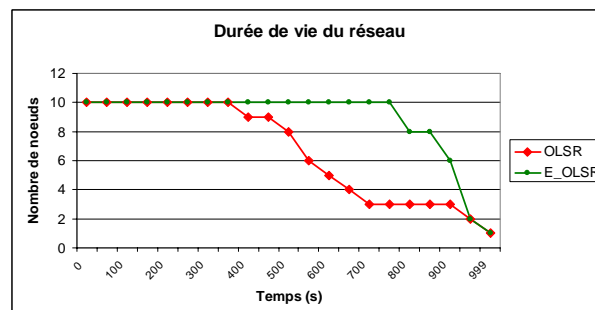


Fig. 7. Durée de vie du réseau

Sur ce graphique on remarque que le nombre de nœuds vivants du réseau OLSR commence à décroître à partir de  $t = 400$  sec, et se stabilise vers  $t = 650$  à cause de la perte de connectivité puis recommence à décroître vers  $t = 900$  sec.

Par contre pour le réseau E\_OLSR le nombre de nœuds vivants reste constant jusqu'à  $t = 800$  sec, où il commence à décroître rapidement. Grâce à l'intégration du poids énergétique  $pe_{\text{}}$  dans la qualité d'un lien (son poids), E\_OLSR favorise les nœuds disposant d'une plus grande énergie résiduelle. À partir de ce graphique, on constate que notre approche a réalisé une augmentation de 22% de la durée de vie moyenne d'un nœud, et par conséquent augmenter aussi la durée de vie de tout le réseau.

### 3 Conclusion

Notre travail a porté sur la conception d'une nouvelle approche pour l'économie de l'énergie dans les réseaux sans fil mobiles au niveau de la couche réseau. L'approche que nous avons proposée coalise deux principes. Le premier principe est le routage orienté puissance, il s'agit de favoriser, pour le routage, les mobiles disposant d'une énergie résiduelle élevée. Le second principe est original, il s'agit d'un rapport entre l'énergie résiduelle d'un émetteur et la distance qui le sépare du voisin récepteur. Ce rapport permet d'équilibrer les mobiles du réseau, ainsi un mobile disposant d'une grande quantité d'énergie et devant transmettre vers un voisin éloigné est équivalent à un mobile disposant d'une quantité d'énergie moyenne et qui doit transmettre sur une distance moyenne ou courte. Après les simulations, nous avons conclu que notre version du protocole OLSR était plus au moins incomplète, vu qu'on s'est fixé l'objectif de ne pas influencer les autres performances du protocole choisis. Quand aux résultats de notre approche, ils s'avèrent très concluants et satisfaisants, avec une économie de l'énergie allant jusqu'à 14% et une augmentation de durée de vie moyenne d'un mobile de 22%.

### References

1. Rong Zheng.: Analysis and Empirical Evaluation of Power Management in Mult-Hop Wireless Networks. Degree of Doctor of Philosophy in Computer Science, University of Illinois at Urbana-Champaign, 2004.
2. Sushant Jain.: Energy Aware Communication in Ad-hoc Networks. Technical Report UW-CSE 03-06-03, Computer Science and Engineering, University of Washington-Seattle.
3. F. INGELREST. : Protocoles localisés de diffusion et économie d'énergie dans les réseaux ad hoc et de capteurs. Thèse de doctorat, Université de LILLE I, France, Juin 2006.
4. J.CARTIGNY. : Contributions à la diffusion dans les réseaux ad hoc. Thèse de doctorat, Université de LILLE I, France, Décembre 2003.
5. L.M. Feeney.: An energy consumption model for performance analysis of routing protocols for mobile ad hoc networks. USA 2001.
6. L.M.Feeney, M.NILSON.: Investigating the energy consumption of a wireless network interface in an ad hoc networking environment. USA 2001.
7. H.BADIS, K.AL AGHA.: Distributed Algorithms for Single and Multiple-Metric Link State QoS Routing. LRI, INRIA, France.
8. T.CLAUSEN, P.JAQUET.: the Optimized Link State Routing protocol. RFC 3626, INRIA, France, Octobre 2003.
9. K.FALL, K.VARADHAN.: The NS manual. Lawrence Berkeley National Laboratory LBL, University of Southern California, USC, 2007.

## Power control issues in mobile ad hoc networks

Benamar KADRI, Mohammed FEHAM

*STIC Lab., Department of Telecommunication,  
University of Tlemcen, Tlemcen, Algeria*

[benamarkadri@yahoo.fr](mailto:benamarkadri@yahoo.fr), [m\\_feham@mail.univ-tlemcen.dz](mailto:m_feham@mail.univ-tlemcen.dz)

**Abstract.** Energy constraint in mobile ad hoc network is one of the most important issues that must be carried out by the creation of new energy aware hardware (battery and network cards), as well as the adaptation of the existed protocols (MAC, routing) to deal with the constraints of energy in mobile ad hoc networks. Thus in this paper we are going to present a new power control scheme based on the number of neighbours in a given neighbourhood, by which we give an improvement to the 802.11 MAC layer in which we include the aspect of transmission power regulation according to the status of the neighbourhood. Simulations given in the last section prove the assumptions and the ideas behind our schemes compared to the existed 802.11 MAC layer.

**Key words:** MANETs, wireless and mobile network, energy, MAC 802.11.

### 1. Introduction

In recent years, the wireless networks have become more and more popular due to low cost, high bandwidth capacities and the facility of deployment, given birth to new standards and hardware [1].

Wireless networks can be classified into two categories, infrastructure network in which a wireless access point play the role of a bridge to an existed wired network, and ad hoc network which is a set of mobile nodes which collaborate between themselves to elaborate a temporary network without the need of any infrastructure or administrative authority, usually composed of a set of handled mobile devices such as cellular phones, PDA, laptops, sensors,...etc [3, 4].

Due to their characteristics such as the topology changing, computing and power constraint, mobile ad hoc network must confront new challenging such as routing, security and quality of service (QoS).

Recently a large volume of research has been concentrated on the issue of energy conservation for wireless networks, since energy conservation is an essential requirement for a wireless ad hoc network system consisting of battery operated nodes.

In ad hoc networks nodes rely on each other to forward packets; therefore each device acts as end host as well as a router to relay packets, since there is no centralized router as in conventional networks, thus most of energy is expended in processing and forwarding packets, which reduces the lifetime of device's battery.

Typically existed battery lifetime is limited to only some hours. Therefore energy conservation is critical in such situation to extend the system and devices lifetime, by defining new issues for existed routing and MAC protocols to save the devices battery power by minimizing the number of operations executed by each node in the network.

### 2. Power control in MANETs

Power control is the mechanism that varies the transmitting power (range) in order to save energy and improve the network capacity. The transmitting power is increased or decreased according to some criteria as the distance between nodes or their movement. The power control has a great importance in mobile wireless network, since it can save the battery power of network devices, and improve the network performance.

## 2.1 Motivation of power control

### 2.1.1 Energy saving

The mobile battery lifetime is an increasingly important issue to manufacturers and consumers, as mobile devices take day after day a great place in our quotidian life in the form of cellular phones, PDAs, laptops..., and usually part of wireless networks such as Personal area networks, wide area networks, sensor networks [4]...etc.

Wireless card used by these devices to allow network connectivity consume the great part of energy of battery to transmit and receive data. The power consumed by the network card is proportional to the transmitting signal strength (power); therefore controlling the strength of the transmitted signal may save the battery power, in the way that the power of the transmitted signal is fixed to the threshold needed to reach the desired destination rather than transmitting with the same signal strength all the time. This issue may save a great amount of energy in mobile ad hoc networks since the mobile devices are usually close to each other [5].

Another issue is considered in power control is the retransmission of data due to collision, according to the specifications of 802.11 MAC [6] layer before any exchange of data, nodes first reserve the medium using CTS/RTS and then transmit data, if the data is received by the destination node, it sends an acknowledgement to achieve this operation otherwise a collision is assumed (Fig.1), and the same operations are repeated until an acknowledgement is received, the repetition of these operations consume a great amount of devices battery power as well as overhearing the network by the number of retransmission. Thus, when considering the issue of power control nodes exchange data within a limited area which decreases the number of collision. And therefore save the device's energy.

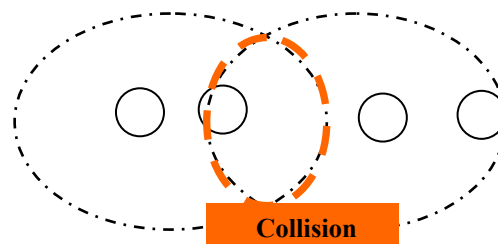


Fig.1. Collision region in wireless networks

### 2.1.2 Capacity improvement

The technique used by the majority of the actual MAC protocols is CSMA/CA [6]. In which any mobile device sense the medium and transmit data if the medium is free, otherwise it fell into a backoff window until the medium becomes free, it also uses RTS/CTS mechanism to avoid the problem of hidden terminal and reserve the medium. As described above this technique consume lot of energy and reduces considerably the network throughput.

Another problem arise is the blockage of all neighbours due to the medium reservation, as shown in Fig. 2; node *A* transmits some data to *B* with a transmission range which exceeds the localisation of *B*, therefore nodes *A*, *B*, *C* can not transmit at the same time according the CSMA/CA scheme until *A* finishes its transmission, therefore any queued data in *C* or *D* is blocked. This problem can be avoided if the transmitting power is fixed to the necessary amount to reach *B*, therefore *C* or *D* can transmit their queued data without waiting for *A* achievement.

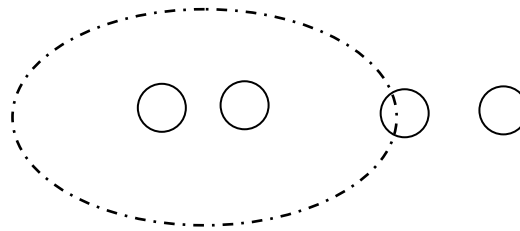


Fig.2 The effect of the transmission range on the network performance.

### 2.1.3 The improvement of security

By nature a wireless network is very open and the transmission range of nodes usually exceeds the area where the network is deployed; therefore any person with the adequate network card and protocols stack can eavesdrop, attack, or use the network services [3]. However when considering a mechanism of power control, the transmission range is fixed to reach only the legitimate nodes and the probability that the signal can be captured by a malicious person is decreased (Fig. 3).

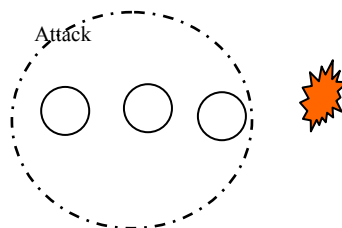


Fig.3 Effect of power control on the security of the network.

## 3. Energy Cost Analysis

Due to the nature of the used medium, when a packet is sent over a wireless link it is captured by all devices in the neighbourhood of the sender, in order to determine either this packet is intended to the corresponding node to be transmitted to the upper layers or not.

The cost of sending a packet begin on the level of the sender, since there is a fixed energy cost in transceiver circuit and a variable cost in the propagation of the signal to reach the receiver which is proportional to the transmission range. However there is another energy cost in receiving a packet by each node in the neighbourhood of the sender which is not negligible. For example, in WaveLAN [7] card with omni-directional antenna, it requires 185 mA for reception and 235 mA for emission.

An additional cost is added by the MAC layer to send a packet which is the power consumed to reserve the medium using signalling messages such as the RTS/CTS requests and the acknowledgement, therefore the total cost of sending a packet is given as:

$$W = ET_x + ER_x + E_m$$

Where  $ET_x$  and  $ER_x$  are the energy consumption for transmission and reception,  $E_m$  is the energy consumption by the MAC layer.

To show best the energy cost in a wireless network let consider the example of Fig. 4 when node *A* wants to transmit a packet to node *B*:

- 1- *A* collaborates with *B* to reserve the medium using RTS/CTS
- 2- If the reservation success then *A* sends the packet to *B*.
- 3- In the same time the messages sent by *A* and *B* are received by *C*, *D* and *F*.



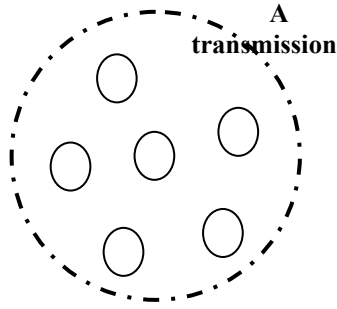


Fig.4 Energy cost analysis.

Therefore, for a successful transmission from *A* to *B* the energy consumed by each node in this neighbourhood is:

Node A:  $W_A = ER_{RTS/CTS} + ET_{RTS/CTS} + ET_{Data} + ER_{Ack}$

Node B:  $W_B = ER_{RTS/CTS} + ET_{RTS/CTS} + ER_{Data} + ET_{Ack}$

Node C:  $W_C = 2ER_{RTS/CTS} + ER_{Data} + ER_{Ack}$

Node D:  $W_D = 2ER_{RTS/CTS} + ER_{Data} + ER_{Ack}$

Node E:  $W_E = 2ER_{RTS/CTS} + ER_{Data} + ER_{Ack}$

Node F:  $W_F = 2ER_{RTS/CTS} + ER_{Data} + ER_{Ack}$

Where:

$ER_{RTS/CTS}$ : the energy consumed for receiving CTS or RTS.

$ET_{RTS/CTS}$ : the energy consumed for sending CTS or RTS.

$ER_{Data}$ : the energy consumed for receiving data.

$ET_{Data}$ : the energy consumed for sending data.

$ER_{Ack}$ : the energy consumed for receiving acknowledgement.

$ET_{Ack}$ : the energy consumed for sending acknowledgement.

Therefore the total energy consumed by a given neighbourhood for a successful transmission of a single data packet is given as:

$W = 10ER_{RTS/CTS} + 2ET_{RTS/CTS} + 5ER_{Data} + ET_{Data} + 5ER_{Ack} + ET_{Ack}$ .

As shown in this equation the energy consumed by the system is very important when the number of receivers' nodes is high, and this issue is very worst if the area where the network is deployed is very small or whenever there is collisions or retransmissions. Therefore, a great amount of the consumed power due to the reception of undesirable data by each node in a given neighbourhood which can be avoided by minimizing the range of transmission to reach only the desired ones.

## 4. Our proposed scheme for power control

### 4.1 Description of our strategy

In our proposed scheme to control power in ad hoc network we have tried to regulate the transmission power according to distance separating the sender and the receiver, in the way that any packet is sent with only the necessary power to reach its destination.

For a given reception threshold power  $P_r$ , the minimum transmits power  $P_t$  at a distance  $d$ , assuming no fading, can be given as:

$$P_r = P_t \times G_r \times G_t \times \frac{\lambda^2}{(4 \times \pi \times d)^2} \dots (1)$$

Where

$P_r$  = received power,

$P_t$  = transmitted power,

$G_t$  = antenna gain of the transmitter,

$G_r$  = antenna gain of the receiver,

$\lambda$  = wavelength,

$d$  = distance.

Therefore when having all information, the distance separating a sender and a receiver can be given as:

$$d = \sqrt{P_t \times G_r \times G_t \times \frac{\lambda^2}{P_r (4 \times \pi)^2}} \dots (2)$$

To compute the distance we must get the gain and the transmitting power of the sender, to do so we have divided our proposed scheme into three phases:

- 1- **Packet capture:** each node captures any transmitted packets and gets from that packet the following information :
  - a. The MAC address of the sender.
  - b. The gain of the antenna of the sender.
  - c. The transmitting power of the sender.
  - d. All this information is saved in a table.
- 2- This information is used after to compute the distance (equation (2)) and therefore the necessary power to reach that node, using the following equation:

$$P_t = TR_r \times \frac{(4 \times \pi \times d)^2}{G_r \times G_t \times \lambda^2}$$

Where  $TR_r$  is the necessary threshold allowing the receiver to correctly receive a packet.

- 3- The computed  $P_t$  is saved in the same table, in order to compute an averaged transmitting power to reach the maximum of neighbours.

$$P_{ar} = \frac{1}{N} \sum P_t$$

To allow this, we have made modification on the original MAC layer packet in order to include in its header the two necessary parameters to compute distance by the receiver which are:

- 1- The gain of the antenna of the sender.
- 2- The transmitting power of sender.

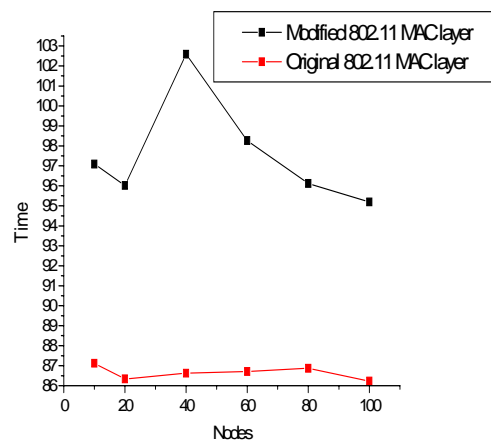
## 5. Simulation results

In order to test the performance of our proposed scheme and test its feasibility we have implemented our scheme in the network simulator NS2 [8], which is the most powerful tool for wireless and wired network simulation since most of the simulation done by researchers are done on NS, we have made modification on the 802.11 MAC and the physical layer in order to implement the specification of our proposed scheme.

To do so we have deployed a set of nodes ranging from 10 to 100 nodes in the area of  $670 \times 670 \text{ m}^2$  in order to test the reaction of our scheme in different cases. Then, we have fixed the number of nodes to 50 and we have varied the pause time, we have also used some CBR (Constant Bit Rate) connections with packet length of 512 Kbytes to emulate traffic over the network, other simulation parameters are listed in table 1.

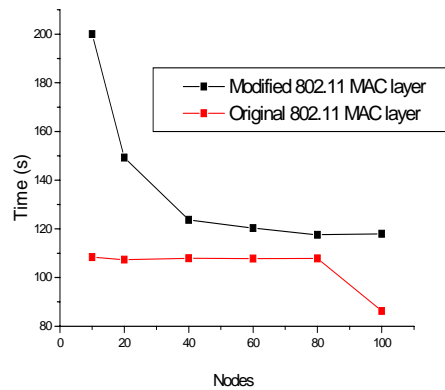
Parameters	Values
Network size	$670 \times 670 \text{ m}^2$
Number of Nodes	10,20,40,60,80,100
Max speed	20 m/s
pause Time	0,50,60,100,150,200s
CBR connections	4
Routing protocol	DSR
Simulation time	200s

Table 1 simulation parameters



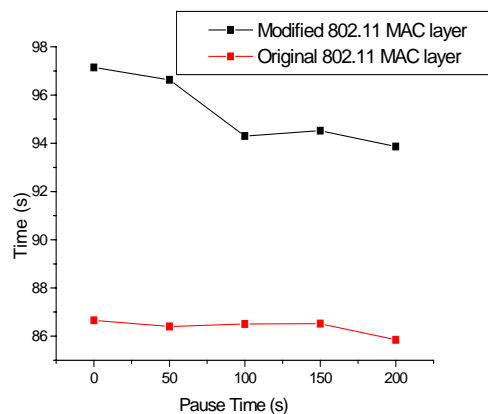
**Fig. 5 The time of the first died node**

In Fig. 5 we have varied the number of nodes from 10 to 100 then we have measured the time at which the first node fails down making the beginning of the system failure. As we can observe our proposed scheme always gives best results and the battery lifetime is extended to more than 10 %.

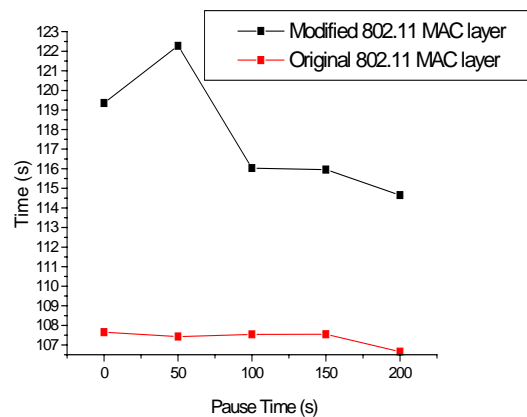


**Fig. 6 System life time**

In Fig. 6 we have measured the system lifetime according to the number of nodes, as shown our improvement to the 802.11 MAC layer always gives best results compared to the original 802.11 MAC layer, and the system lifetime is expended to more 40 % in some situation, and more than 10 % in the majority of cases.



**Fig. 7 The time of the first died node**



**Fig. 8 System life time**

In Fig. 7 and Fig. 8 we have fixed the number of nodes to 50 nodes and we have varied the pause time from 0 s to 200 s (static network); and we have measured the time of the first died node and the system lifetime. As we can observe our proposed scheme always gives best results and the battery and the system lifetime is extended to more than 20%.

## 6. Conclusion

As presented above, power control in wireless network is an important issue since it can solve lot of problems in addition to the conservation of the mobile nodes battery, such as the security of the network and the improvement of the quality of service.

Our modifications on the network stack are done on the 802.11 MAC layer, since this layer is direct contact with the physical layer responsible for data transmission on the medium.

In our scheme the transmission power for each node is computed in the MAC, and transmitted to the physical layer to regulate the power for transmission.

As shown in the last section, simulation results show that our proposed scheme is energy efficient since it conserves the battery power of node and extends the system lifetime to more than 10 % which is very good compared to the original MAC layer and other MAC layers.

In future works we will try to implement our scheme in real situation with the existence of the fading and noise and we will try than to implement it for other kinds of network as sensor and personal networks

## REFERENCE

1. Houda Labiod, Hossam Afifi. De Bluetooth a WIFI sécurité, qualité de service et aspects pratiques, hermes-science, 2004.
2. B.KADRI, A.M'HAMED, M. FEHAM. "Secured Clustering Algorithm for Mobile Ad hoc Networks". International Journal of Computer Science and Network Security, March 2007.
3. B.KADRI, A.M'HAMED, M. FEHAM. "A new management scheme of cluster based PKI for ad hoc networks using multi-signature". IEEE Global Information Infrastructure Symposium, 2007.
4. I.F. Akyildiz, W. Su\*, Y. Sankarasubramaniam, E. Cayirci, "Wireless sensor networks: a survey", Computer Networks 38 (2002) 393–422.
5. W. Li, Z. Bao – yu, "Study on Cross – Layer Design and Power Conservation in Ad hoc Network", IEEE PDCAT'2003, 27-29 Aug. 2003 Pages:324 – 328.
5. IEEE Computer Society LAN MAN Standards Committee. "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications" IEEE Std 802.11-1997.
7. ORINOCO. Orinoco wavelan radio. <http://www.orinoco.com> , 2000.
8. The Network simulator The Network Simulator ns-2. Project web page available at <http://www.isi.edu/nsnam/ns/>.

## Protocole de Reprise Efficace pour Applications Ubiquitaires dans les Réseaux Mobiles à Infrastructure.

Nabila Labraoui <sup>1</sup>, Makhoulf Aliouat <sup>2</sup>

Département Informatique, UFA, Sétif, Algérie

<sup>1</sup> labraouinabila@yahoo.fr

<sup>2</sup> aliouat\_m@yahoo.fr

**Résumé :** De récents progrès technologiques remarquables tant dans le domaine des terminaux portables que celui des communications sans fil, conjugués à la baisse des prix des composants et à une forte demande des services nomades ont conduit à un développement notable des réseaux mobiles. Néanmoins, disposer pleinement de ces environnements nouveaux nécessite de relever le défi imposé par leurs caractéristiques physiques contraignantes. La tolérance aux fautes, un des attributs majeurs des systèmes distribués fixes, doit s'adapter au contexte mobile afin d'offrir aux applications ubiquitaires une continuité de service satisfaisante malgré un environnement ambiant précaire soumis à un taux d'occurrence d'erreurs important. La technique d'établissement de points de reprise (checkpointing), éventuellement associée à une sauvegarde des messages (message logging), permet de reprendre (à moindre coût) une application erronée, lui garantissant ainsi une délivrance des résultats conformes aux délais attendus (spécifiés). Dans cet article, nous proposons un protocole d'établissement de points de reprise global cohérent dédié aux applications distribuées s'exécutant dans un réseau mobile à infrastructure. Basé sur une coordination efficace des sites mobiles causalement dépendants, ce protocole assure une sauvegarde d'un point de reprise global cohérent exempt de points de reprise locaux inutiles (minimisation de l'overhead espace/temps) et contribue de manière significative à la préservation et la conservation de l'énergie des terminaux portables. Une évaluation des performances et comparaison avec d'autres protocoles connus indiquent que notre proposition occupe une place appréciable.

**Mots clés :** Applications distribuées mobiles, réseaux mobiles à infrastructure, tolérance aux fautes, checkpointing, état global cohérent.

### 1. Introduction

De récentes avancées technologiques tant dans le domaine des terminaux portables (baisse des coûts, miniaturisation, fonctionnalités diverses etc.) que dans celui des réseaux mobiles ont conduit à une forte activité de recherche dans ce qu'il convient d'appeler paradigme des environnements informationnels mobiles. Ces derniers ont été les précurseurs d'un nouvel axe prometteur d'investigation : l'informatique nomade (mobile ou ubiquitaire) soustraite aux contraintes du lieu et du temps. Le déploiement de ces environnements mobiles introduit de nouvelles problématiques issues de leurs spécificités intrinsèques physiques et environnementales contraignantes.

Les réseaux mobiles se caractérisent par la mobilité des nœuds, leurs fréquentes déconnexions, la limitation de leurs ressources telles que : bande passante faible, durée de vie des batteries limitée, capacités mémoire et de traitement faibles et surtout une grande vulnérabilité aux défaillances et aux malveillances. Le taux d'occurrences des erreurs y est relativement important, et mener à terme une application distribuée ubiquitaire relève d'un défi certain. Compte tenu de leurs spécificités propres, les réseaux mobiles ne peuvent directement tirer parti des techniques de tolérance aux fautes déjà éprouvées dans les réseaux fixes. Il est donc crucial soit de les adapter au mieux à ce nouveau contexte, soit d'en concevoir de nouvelles spécialement dédiées. Ainsi, un usager mobile souhaiterait se déplacer librement tout en s'assurant une continuité de service malgré les erreurs pouvant affecter son application. Autrement dit, doter cette dernière de capacité à tolérer les erreurs par le biais de recouvrement arrière ou reprise après erreur.

Le recouvrement arrière est un mécanisme suffisamment général et efficace (moindre coût) pour tolérer les erreurs en cours d'exécution et assurer aux applications une délivrance des résultats attendus conforme à l'échéance spécifiée. Il est fondé sur la sauvegarde régulière de l'état global d'une application distribuée ou d'un système, appelée aussi point de reprise (ou checkpoint), à partir duquel les processus peuvent être relancés en cas d'échec. Ce point de reprise permet d'éviter de ré-exécuter l'application depuis son début, épargnant ainsi une perte de temps considérable et inacceptable (surtout s'il s'agit d'une application à longue durée d'exécution).

Concevoir un protocole de reprise adéquat, signifie détecter les dépendances causales entre processus communiquant pour construire un checkpoint *cohérent*. Cette cohérence est mise en défaut si un checkpoint global contient au moins un message orphelin [7], c'est-à-dire un message émis dont l'évènement d'émission a eu lieu après un enregistrement d'état local (appartenant à un état global) et l'évènement de réception s'est produit avant un enregistrement d'état local. L'état global au coût minimal souhaité, est celui constitué d'états locaux (un par processus) les plus récemment établis. Il est appelé ligne de recouvrement et minimise le retour arrière en cas de reprise après erreur.

Dans cet article, nous proposons un nouveau protocole d'établissement de point de reprise (checkpointing) basé sur la coordination des processus causalement dépendants. Il engage un nombre minimum de processus dans la procédure de checkpointing (conséquence : overhead temps/espace minimal). Il est efficace, contribue de manière significative à préserver l'économie de l'énergie du terminal portable et assure un état global cohérent et une reprise rapide. Et enfin, il prend en charge la mobilité et la déconnexion volontaires des sites mobiles.

Le reste de l'article est structuré comme suit: la section 2 présente des travaux similaires significatifs de quelques algorithmes de checkpointing en environnement mobile. Dans la section 3, nous faisons une critique de deux algorithmes coordonnés connus. Dans la section 4, nous présentons en détail notre proposition de protocole de reprise. Dans la section 5, nous évaluons les performances de notre protocole et enfin nous terminons ce travail par une conclusion dans la section 6.

## 2. Travaux similaires

Acharya et Badrinath, ont été les premiers à proposer dans [1] un protocole de checkpointing dédié au contexte mobile. Ce protocole se compose de deux parties : une partie exécutée par chaque site mobile pour sauvegarder de manière asynchrone son point de reprise, et l'autre exécutée par la station de base, pour le compte d'un site mobile, afin de déterminer une ligne de recouvrement cohérente. Prakash et Singhal ont proposé dans [2] un algorithme coordonné non bloquant, qui force un nombre minimum de nœud impliqués. Il est réalisé en deux phases : La phase *Lazy* est quasi coordonnée, et la phase *aggressive* est coordonnée. Deux ans après, les mêmes auteurs, prouvent que cet algorithme génère des états d'incohérence. Ils prouvent ensuite dans [3] qu'il n'existe pas d'algorithme de checkpointing qui soit à la fois non bloquant et minimal. Neves et Fuchs proposent dans [4] un protocole de Checkpointing basé sur le temps physique. Les sites sauvegardent périodiquement leurs états à l'expiration d'un temporisateur local. Dans [5], Cao et Singhal présentent le concept de points de reprise *Mutables* permettant d'assurer la cohérence du checkpoint obtenu mais génère un overhead conséquent contribuant à l'épuisement de la batterie. Dans [6] Higaki et al. ont proposé un algorithme Hybride, combinant la stratégie coordonnée pour la sauvegarde d'un état global cohérent pour les sites fixes, et la stratégie non coordonnée pour la sauvegarde de l'état global des sites mobiles. Cette combinaison attrayante induit néanmoins un overhead non négligeable. Dans [7] Lin et Dow proposent un nouvel algorithme hybride combinant l'approche coordonnée des sites fixes et l'approche induite par les communications, pour la sauvegarde des points de reprises des sites mobiles. Sapna et al. ont proposé dans [9] un algorithme basé sur le déplacement des sites mobiles et combinant l'approche non coordonnée à la journalisation optimiste des messages. Gupta et Rahimi ont proposé dans [10] un algorithme de recouvrement efficace et simple basé sur l'approche non coordonnée. Lors d'un échec, un état global cohérent est calculé. Gupta et al. ont présenté dans [11] un algorithme de checkpointing coordonné, minimal en utilisant le blocage du traitement des messages qui risquent d'être orphelins.

## 3. Constat et critique

Pour éviter les messages orphelins (cause de l'incohérence des checkpoints), deux méthodes sont utilisées: la première [5] consiste à faire prendre à chaque processus un checkpoint additionnel avant le traitement d'un message qui risque d'être orphelin. Ce checkpoint est soit rendu permanent ou annulé dans le cas où il n'est plus nécessaire. Cette attitude s'avère coûteuse en énergie découlant du gaspillage temps/espace qui en résulte sachant bien que le contexte mobile en est très sensible. Dans la seconde méthode, on impose un blocage temporaire rigide d'un processus si un message reçu est susceptible d'être orphelin [11].

Afin de remédier aux deux problèmes cités ci-dessus, nous proposons un algorithme qui minimise la prise de checkpoints inutiles et allège le blocage rigide des processus en utilisant le *blocage partiel* et la notion de *checkpoint retardé*.

## 4. Proposition d'un protocole de checkpointing

### 4.1 Idée de base

L'idée principale est de bloquer le traitement des messages qui risque d'être orphelins. Ce blocage n'est pas effectué systématiquement et peut être minimisé tout en préservant la cohérence de l'état global. Par exemple, si le processus a déjà pris un checkpoint correspondant au même initiateur, alors le message reçu est normalement traité. Un message peut devenir orphelin dans les deux cas suivants : - Lorsqu'un processus envoie un message, après avoir pris un point de reprise, - Lorsqu'un processus reçoit un message avant de prendre un point de reprise. En utilisant ces informations, et la notion de *points de reprise retardés*, on arrive à minimiser le blocage et préserver la cohérence. Ce checkpoint local *extra* va permettre d'éviter les messages orphelins. Dans l'exemple de la figure 1, si P1 prend un autre point de reprise après l'envoi de m1, alors m1 ne devient plus orphelin. Ce point de reprise est appelé "checkpoint retardé". Il est pris par un processus dans la deuxième phase de la procédure de checkpointing lorsqu'il reçoit une validation de l'initiateur afin de rendre son checkpoint permanent. Cela est effectué seulement s'il a envoyé au moins un message après avoir pris un checkpoint dans la première phase.

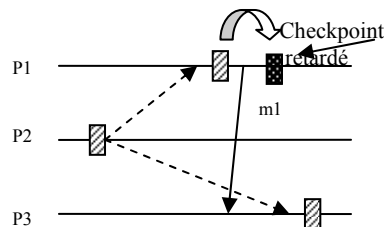


Fig 1: Principe du checkpoint retardé.

Dans la deuxième phase, le processus P1 rend ce checkpoint retardé permanent à la place du précédent. Cette approche est différente de celle des checkpoints mutables, car ce point de reprise additionnel n'est jamais effacé, par conséquent, la consommation d'énergie qui est dépensée n'est pas gaspillée. Cependant, la prise du checkpoint retardé n'est pas possible dans toutes les situations. Soit l'exemple suivant : Supposons que P2 reçoive le message m2 de P3 (fig.2.) avant d'envoyer m1, si à ce moment P2 prend un checkpoint retardé, m2 devient orphelin. Dans ce cas, on utilise une indication de blocage de traitement d'un message « *receiver-based blocking* », implémentée par un bit dans l'entête du message : *block\_bit*. Le *block\_bit* est mis à 1 si l'expéditeur veut que le récepteur du message bloque le traitement de ce dernier.

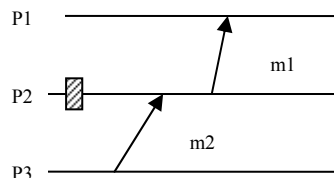


Fig2 : Un cas particulier où le checkpoint retardé ne peut être pris.

Dans l'exemple précédent P2 doit envoyer le message m1 avec le *block\_bit* égal à 1. Ceci indique que P1 doit sauvegarder m1 dans une file et bloquer son traitement pour éviter plus tard à P1 de prendre un checkpoint faisant de m1 un message orphelin. Donc, les trois conditions qui doivent être réunies pour bloquer le calcul d'un processus sont : - l'expéditeur de m a pris un checkpoint, - Avant l'envoi de m, l'expéditeur a reçu un autre message d'un autre processus, - Le récepteur de m n'a pas pris de checkpoint.

Ce blocage est appelé : blocage partiel (*partial blocking*). Durant le blocage, si un processus reçoit une requête de checkpointing, il prend alors un point de reprise et commence à traiter les messages bloqués dans la file dans l'ordre de leur arrivée. S'il ne reçoit pas de requête, il attend un message de fin de blocage « *blocking\_clear\_message* » pour arrêter le blocage des messages et procéder à leur traitement.

Ce *blocking\_clear\_message* est envoyé dans la deuxième phase par le processus qui a envoyé le message qui devait être bloqué. Ce message annonce la fin du checkpointing et le traitement des messages ne peut plus causer d'incohérence. Puisque l'algorithme est destiné à des applications non déterministes, alors le blocage des messages qui risquent d'être orphelins n'est qu'une manière de différer ou de retarder leur traitement afin de maintenir la cohérence de l'état global.

## 4.2 Modèle du système

On considère un système de  $N$  processus  $P_i$   $i \in [1, N]$ , répartis sur un réseau mobile à infrastructure composé de stations fixes appelées *MSS* (Mobile Support Station) et de terminaux mobiles *MH* (Mobile Host). Une *MSS<sub>i</sub>* est responsable des *MHs* situés actuellement dans sa cellule *Cell<sub>i</sub>*. Une cellule est un secteur géographique qui regroupe plusieurs *MHs* à portée de communication d'une *MSS*. La communication entre un *MH<sub>j</sub>* et son *MSS<sub>i</sub>* est sans fil avec une technique *FIFO* d'émission-réception des messages. Un *MH<sub>j</sub>* ne peut communiquer qu'avec son *MSS<sub>i</sub>* qui le gère. La communication entre deux sites mobiles *MH<sub>i</sub>* et *MH<sub>j</sub>* passe obligatoirement par leurs stations de base responsables *MSS<sub>i</sub>* et *MSS<sub>j</sub>*. Les stations de base sont interconnectées par un réseau fixe et communiquent par l'envoi de messages avec une stratégie *FIFO*. Cependant, on suppose que le taux de tolérances aux fautes et la capacité de calcul pour les stations de base et le réseau fixe sont très élevés.

## 4.3 Prise en compte de la mobilité

Pour traiter les déplacements ou hand off, un site mobile doit informer lui même les stations de base concernées de tous ses déplacements. Ainsi, quand un site mobile se déplace d'une cellule à une autre, il informe l'ancienne et la nouvelle station de base de ce déplacement. Cependant, des mesures doivent être prises pour s'assurer que les informations de dépendance, le numéro du dernier point de reprise et éventuellement la liste des processus bloqués soit toujours au niveau de la station de base courante (Station de base de la cellule actuellement visitée par le site mobile).

## 4.4 Prise en compte des déconnexions

La déconnexion est aussi un autre problème qui affecte les performances du checkpointing. Nous prenons en compte les déconnexions volontaires, tandis que celles qui ne le sont pas sont considérées comme des fautes :

1- avant la déconnexion, le site mobile *MH<sub>i</sub>* prend un point de reprise et le sauvegarde dans la station de base courante *MSS<sub>i</sub>* comme un checkpoint de déconnexion : (*DeconnexionCkpt*). Ainsi, la *MSS<sub>i</sub>* répond avec ce point de reprise si une demande de checkpointing arrive pour *MH<sub>i</sub>* pendant l'intervalle de déconnexion. Il sauvegarde éventuellement la liste des processus bloqués (*Blocking\_Set<sub>i</sub>*), au niveau de *MSS<sub>i</sub>*, et ce, pour éviter le blocage indéfini des processus si *MH<sub>i</sub>* se déconnecte avant la fin de la procédure de checkpointing, (i.e. avant d'envoyer la requête de déblocage). Dans ce cas, c'est *MSS<sub>i</sub>* qui se charge de débloquer les processus à sa place.

2- A la reconnexion, le *MH<sub>i</sub>* exécute une routine de reconnexion en envoyant l'identité de la dernière station de base visitée *MSS<sub>i</sub>*, à la nouvelle station de base courante *MSS<sub>j</sub>*. *MSS<sub>j</sub>* pourra ainsi récupérer le numéro du dernier point de reprise de *MH<sub>i</sub>*, son vecteur de dépendances et le journal des messages arrivés au profit de *MH<sub>i</sub>* pendant sa déconnexion et éventuellement la liste des processus bloqués. Voir [8], [12].

## 4.5 Description de l'algorithme

L'algorithme proposé utilise l'approche coordonnée minimale et se déroule en deux phases :

**Première phase:** Lorsqu'un site mobile *MH<sub>init</sub>* initie l'algorithme, *MH<sub>init</sub>* enregistre localement un point de reprise provisoire et envoie une requête de coordination à sa station de base *MSS<sub>init</sub>*. (Cette dernière, va se charger de toute la procédure de checkpointing). A la réception de la requête, la *MSS<sub>init</sub>* détermine les identificateurs de l'ensemble *Direct\_Dep* (sites mobiles dépendants directs de *MH<sub>init</sub>*) et diffuse ensuite une requête de propagation de Checkpointing à toutes les stations de base (y comprise elle-même). Cette requête encapsule l'ensemble *Direct\_Dep* déterminé précédemment. Après l'initiation



et pendant toute l'exécution de l'algorithme, si une station de base  $MSS_i$  reçoit une requête *Propchp*, alors  $MSS_i$  diffuse à son tour cette requête aux sites  $MH_i \subset Cell_i$  concernés par le *checkpointing*.  $MSS_i$  détermine également les identificateurs des sites mobiles qui sont dépendants directement des sites mobiles indiqués dans l'ensemble *Direct\_Dep* (dépendants indirectement du site mobile initiateur  $MH_{init}$ ). Si l'ensemble ainsi déterminé est vide, alors  $MSS_i$  envoie un acquittement à la station initiatrice  $MSS_{init}$ , sinon elle diffuse cet ensemble à toutes les autres stations de base, et ainsi de suite.

**Deuxième phase :** Quand la station de base initiatrice  $MSS_{init}$  reçoit une réponse de toutes les  $MSS_j$ , alors elle diffuse une requête de validation *Commit* (des points de reprise provisoires) vers toutes les stations de base. Chaque station de base, ayant reçu une requête de validation, diffuse à son tour cette requête dans sa propre cellule. Ainsi, chaque site mobile ayant reçu la requête *Commit*, envoie son point de reprise provisoire à sa  $MSS$  pour le transformer en point de reprise permanent. De même, chaque  $MSS_j$  envoie un message *blockClear* à tous les sites  $MH_j$  ( $1 \leq j \leq k$ ) de sa cellule  $Cell_j$  dont le traitement de certains messages a été bloqué, afin que ces derniers puissent être traités.

### Structure de données

Pour chaque site mobile  $MH_i$

<b>ddv<sub>i</sub></b> : Array [1..n] of Bit ;	// vecteur des dépendances directes
<b>State<sub>i</sub></b> : bit ;	// état du processus, State <sub>i</sub> =1 alors $MH_i$ a pris un checkpoint
<b>Sent<sub>i</sub></b> : bit ;	// si Sent <sub>i</sub> =1 => $MH_i$ doit prendre un checkpoint retardé
<b>Block<sub>i</sub></b> : bit ;	// si Block <sub>i</sub> =1 => $MH_i$ doit bloquer le processus à qui il va envoyer un msg
<b>Ckpt<sub>i</sub></b> : integer;	// numéro de séquence du dernier point de reprise
<b>Blocking_Set</b> : Set of identifiant	//ID des process bloqués par $MH_i$

Pour chaque station de base  $MSS_i$ :

<b>Coordinator</b> : Identifiant;	// ID de la MSS coordinatrice
<b>Finish</b> : real ;	// sert à la terminaison de l'algorithme (protocole de Chang)
<b>Cell<sub>i</sub>, Disconnect</b> : Set of Identifiers;	// ID des sites de la cellule et des sites déconnectés
<b>Direct_Dep, Trans_Dep</b> : Set of Identifiers ;	// ID des sites dépendants directs et indirects
<b>Total_dep</b> : Set of Identifiers;	// ID de tous les sites invoqués dans le checkpointing
<b>TBlock_Set</b> : array [1..n] of bit init 0;	// Tblock_Set[i]=1 => $MH_i$ a été bloqué ;
<b>SendClear</b> : array [1..n] of bit init 0 ;	//SendClear[i]=1 => $MH_i$ a déjà reçu un blockclear msg

### 4.6 Pseudo code de l'algorithme

#### 1- Lorsque $MH_i$ envoie un message de calcul à $MH_j$

```

If (Statei = 1) Then
  If (Blocki = 1) Then
    Block_bit = 1
    Blocking_Seti := blocking_Seti  $\cup$   $MH_j$  ;
  Else Block_bit = 0
    Senti = 1 ; EndIf
  Else Block_bit = 0; EndIf ;
SendMessage( $MH_i, MH_j$ , block_bit, statei, Message) to  $MH_j$  ;

```

#### 2- Lorsque $MH_i$ reçoit un message de calcul de $MH_j$

```

receiveMessage ( $MH_j, MH_i$ , m_block_bit, m_state, Message)
If m_state = 1 Then
  If ( m_block_bit = 0 ) Then
    If (Statei=1) Then Blocki=1; Senti=0; Process the message
    Else Process the message EndIf
    Else If (Statei = 1) Then Blocki=1; Senti=0;
      Process the message
    Else put message in blocking-queue
    EndIf ; EndIf
  Else process the message ;
    ddvi[j] = 1 ; EndIf ;

```

**3- Lorsque  $MSS_i$  reçoit un message de calcul de  $MH_{exp}$  vers  $MH_{des}$  :**

```
ReceiveMessage (MHexp, MHdes, block_bit, stateexp, Message)
If MHdes ∈ Celli Then
    SendMessage(MHexp, MHdes, block_bit, stateexp, Message) to MHdes
    If statedes = 0 Then          ddvdes [exp] :=1 EndIf
    Else MSSdes := search (MHdes) ;
    SendMessage(MHexp, MSSdes, MHdes, blockbit, stateexp, Message) to MSSdes
EndIf ;
```

**4- Lorsque  $MSS_i$  reçoit un message de calcul de  $MSS_j$  :**

```
ReceiveMessage(MHexp, MSSdes, MHdes, block_bit, stateexp, Message)
SendMessage(MHexp, MHdes, block_bit, stateexp, Message) to MHdes
If stateexp = 0 Then          ddvdes [exp] :=1 EndIf;
If block_bit =1 then TBlockSet [exp] :=1 EndIf;
```

• **Phase 1 :**

**5- Lorsque le site mobile  $MH_{init}$  initialise une procédure de checkpointing**

```
Take a Tentative Checkpoint; Stateinit = 1 ;
If Empty ( ddvinit [ ] )
    Then SendCkpt (MHi, TentativeCkpti) to local MSS;
    Else SendCoordReq(MHinit) to local MSS ; EndIf ;
```

**6- Lorsque la  $MSS_{init}$  reçoit une requête de coordination de  $MH_{init}$  :**

```
ReceiveCoordReq (m_MHinit )
First_req:=1; Finish:= 0;
Direct_Dep:= {MHi / ddvinit [i] := 1 ∀ i ∈ [1, n]};
Total_dep := Direct_Dep ;
For All MSSj Do SendPropChp( MSSinit , Direct_Dep, 1/m, first_req ) EndDo;
```

**7- Lorsque  $MSS_i$  reçoit une requête de propagation de checkpointing de  $MSS_{init}$  ou de  $MSS_j$  :**

```
ReceivePropChp (m_MSSinit, Set_Dep, m_weight, m_first_req )
If m_first_req =1 Then
    coordinator :=m_MSSinit ; First_req := 0 EndIf ;
    Weighti := m_weight ;
    Direct_Dep:= {MHi / MHi ∈ Set_Dep ∧ MHi ∈ Celli ∧ MHi ∉ Total_Dep ∧ MHi ∉ Disconnect ∧
    MHi <> MHinit} ;
    Total_Dep := Total_Dep ∪ Set_Dep ;
    For all (MHi ∈ Direct_Dep) Do SendChpReq(MHi) ; EndDo
    Trans_Dep:= {MHi / (ddvk [i] :=1) ∧ (MHk ∈ Direct_Dep) ∧ (MHi ∉ Total_Dep) ∀ i, k ∈ [1
    , n]}};
    If (Trans_Dep ≠ ∅) Then
        For All MSSj Do SendPropChp(MSSinit , Trans_dep, weighti/m , first_req )
EndDo ;
    Else SendReply(Coordinator, weighti) to Coordinator ; EndIf ;
```

**8- Lorsque  $MH_i$  reçoit une requête de checkpointing de la part de la station de base  $MSS_j$**

```
receiveChpReq ()
Statei = 1; Take the tentative checkpoint;
```

• **Phase 2**

**9- lorsque le coordinateur  $MSS_{init}$  reçoit une réponse**

```
ReceptReply( m_weight); Finish := Finish + m_weight ;
If (Finish = 1) Then For all MSSj Do SendCommit() EndDo EndIf ;
```

**10- Lorsque la station de base  $MSS_j$  reçoit une requête de validation**

```
ReceiveCommit( )
For all (MHi ∈ Celli ∧ MHi ∈ Total_Dep ∧ MHi ∉ Disconnect) Do SendCommit ( )
EndDo;
```

```

For all  $MH_i / TBlock\_Set[i]=1 \wedge SendClear[i]=0 \wedge MH_i \in Total\_Dep \wedge MH_i \notin Deconnect$  Do
    SendBlockClear (); Sendclear[i] :=1 ; EndDo ;
    
```

#### 11- Lorsque la MSS reçoit une requête de sauvegarde d'un checkpoint provisoire de $MH_i$ :

```

ReceiveCkpt( $MH_i, TentativeCkpt_i$ ) ; PermananCkpti:=CkptProvisoire
SaveCkpt ( $MH_i, Ckpt_i, PermanentCkpt_i$ ) ;
    
```

#### 12- Lorsque la MSS reçoit la dernière requête de sauvegarde d'un checkpoint provisoire :

```

For all ( $TentativeCkpt_i$ ) Do PermanentCkpti:= TentativeCkpti ;
    SaveCkpt ( $MH_i, Ckpt_i, PermanentCkpt_i$ ) ; EndDo ;
If ( $\exists DeconnexionCkpt_i \wedge (\exists CkptProvisoire_i) \wedge (MH_i \in Total\_dep)$ ) Then
    CkptPermanenti:= CkptDéconnexioni ; Ckpti := Ckpti + 1 ;
    SaveCkpt ( $MH_i, Ckpt_i, PermanentCkpt_i$ ) ;
    If ( $Blocking\_Set_i \neq \emptyset$ ) Then
        For all  $MH_j / (Blocking\_Set_i[j]=1 \wedge SendClear[j]=0 \wedge MH_j \in Total\_Dep \wedge MH_j \notin Deconnect)$  Do
            SendBlockClear (); Sendclear[j]:=1 ; EndDo ;
        EndIf ; EndIf ;
    For All_ ( $MH_i \in Total\_Dep$ ) Do DeleteCkpt ( $MH_i, Ckpt_i - 1, PermanentCkpt_i$ ) ;
EndDo ;
    
```

#### 13- lorsque le Site $MH_i$ reçoit un commit message

```

ReceptCommit()
If ( $block_i=1$ ) Then SendBlockClear( $Blocking\_set_i$ ) to local MSS EndIf ;
If  $\exists TentativeCkpt_i$  Then
    SendCkpt ( $MH_i, TentativeCkpt_i$ ) to local MSS ; EndIf
Else //blocki =0
    If ( $sent_i=1$ ) Then Take DelayCkpt ; TentativeCkpti:= DelayCkpti ;
        SendCkpt ( $MH_i, TentativeCkpt_i$ ) to local MSS ;
    Else SendCkpt ( $MH_i, TentativeCkpt_i$ ) to local MSS ; EndIf ; EndIf ;
    
```

#### 14- lorsque un $MH_j$ reçoit un blockClear message

```

ReceiveBlockClear ();
Process the message in the blocking queue;
    
```

## 5. Evaluation

Nous présentons dans cette section une évaluation des performances de notre algorithme, ainsi qu'une comparaison par rapport à deux autres algorithmes connus étudiés de la même classe.

Nous utilisons les notations suivantes pour évaluer les performances de l'algorithme :

$C_{DiffMob}$  : Coût de diffusion d'un message de contrôle dans le réseau mobile.

$N$  : Nombre total de sites mobiles dans le système.

$M$  : Nombre total des stations de base dans le système.

$N_{Dep}$  : le nombre exact de sites mobiles dépendants directement ou indirectement du site mobile initiateur ;

$$1 \leq N_{Dep} \leq N.$$

$N_{Min}$  : Nombre de sites mobiles établissant des points de reprise :

$$1 \leq N_{Min} \leq N_{Dep} < N.$$

### 5.1 Performances de notre algorithme

a) **Nombre de points de reprise créés pendant l'exécution de l'algorithme** : seuls les sites mobiles dépendants directement ou indirectement du site mobile initiateur sont contraints d'établir des points de reprise. Donc, un nombre minimum ( $\approx N_{Dep}$ ) de sites mobiles doivent prendre des checkpoints.

**b) Travail refait lors de la reprise :** l'utilisation du concept du checkpoint retardé évite, lors de la reprise, de refaire le travail déjà effectué (i.e. l'émission, la réception et le traitement des messages). La quantité de travail à refaire est donc portée au minimum et le recouvrement s'opère rapidement.

**c) Coût en messages de contrôle :** pendant la première phase, le site mobile établissant un point de reprise provisoire n'a besoin que d'un seul message de contrôle : requête de Checkpointing envoyée à sa station de base, d'où le nombre de messages de contrôle égale à  $\approx N_{Dep}$ , et le coût est égal à  $\approx N_{Dep} * C_{DifMob}$ .

Pendant la deuxième phase, chaque station de base diffuse, dans sa cellule, une requête de validation pour que tout site mobile ayant précédemment pris un point de reprise provisoire, envoie ce dernier à sa station de base, et éventuellement un message de déblocage pour les sites qui ont bloqué le traitement de certains de leurs messages reçus. Le nombre de messages de contrôle pendant la deuxième phase est égal à  $\approx N_{Dep}$ , et le coût égale à  $\approx N_{Dep} * C_{DifMob}$ . Donc, le coût total de diffusion des messages de contrôle est égal à :  $C_{DifMob} * (2 * N_{Dep})$ .

**d) Blocage du traitement des messages :** le blocage du traitement des messages susceptibles d'être orphelins n'est pas systématique. Il est allégé en utilisant le blocage partiel. Ce blocage est différent de celui exercé par d'autres algorithmes bloquants, soit en interdisant l'envoi des messages ou en interdisant l'envoi et la réception de messages [3]. Dans notre algorithme, le blocage ne concerne que le traitement de certains messages qui risquent d'être orphelins, alors que le processus concerné peut émettre et recevoir sans aucune contrainte.

## 5.2 Comparaison avec d'autres algorithmes :

Nous comparons notre algorithme avec deux algorithmes de Checkpointing coordonnés significatifs :

(1) L'algorithme de Cao et Al.: *Mutable Checkpoints* [5] et

(2) L'algorithme de Gupta et Al : *A Non-intrusive Snapshot Collection Protocol for Distributed Mobile System*, proposé en 2007 [11].

**Tab1** : comparaison des algorithmes.

Les algorithmes	Nombre de points de reprise	Etendue du travail refait	Coût de synchronisation	Blocage des messages
1	$N_{Min}$	Beaucoup	$C_{DifMob} * (4 * N_{Min} + 1)$	Aucun
2	$N_{Dep}$	Beaucoup	$C_{DifMob} * (2 * N)$	Systématique
Notre algorithme	$N_{Dep}$	Très peu	$C_{DifMob} * (2 * N_{Dep})$ .	Minimisé

## 6. Conclusion

Après analyse minutieuse de différents algorithmes significatifs de reprise dédiés aux réseaux mobiles avec infrastructure nous en proposons un nouveau dont la contribution réside dans : - Optimisation du nombre de points de reprise établis : seuls les processus causalement dépendants sont concernés par l'opération de checkpointing, ce qui implique une génération d'overhead espace/temps minimum, - Traitement de la charge du processus de checkpointing dans les stations de bases : les stations mobiles ainsi déchargées s'occuperont du traitement effectif des processus des applications mobiles distribuées, ce qui résulte d'une économie substantielle de la bande passante et d'énergie au niveau des stations mobiles (optimisation des émissions de messages de contrôle), - Gaspillage minimum du travail défait par le biais des points de reprise retardés et blocage minimum des processus d'application, d'où ralentissement réduit du temps de réponse des application et respect des échéances de délivrance des services. L'analyse des performances de notre proposition vis-à-vis des algorithmes de même type et relativement ou coût global généré, montre que notre proposition se situe en bonne place. Toutefois, une simulation du comportement du protocole (en perspective) permettrait de fixer au mieux certains paramètres tels que : le Temps de Convergence du protocole TC, l'influence de la mobilité sur TC, le taux de messages erronés et perdus.

## Bibliographie

- [1] A. Acharya and B. R. Badrinath. « Checkpointing Distributed Applications on Mobile Computers ». the Third Intl. Conf. on Parallel and Distributed Information Systems. Sep 1994.
- [2] R. Prakash and M. Singhal. « Low-Cost Checkpointing and Failure Recovery in Mobile Computing Systems » IEEE Trans. Parallel and Distributed Systems, PP. 344-351, Oct 1996.
- [3] G. Cao and M. Singhal. « On the Impossibility of Min-Process Non-Blocking Checkpointing and an Efficient Checkpointing Algorithm for Mobile Computing Systems » Proc 27 th Intl Conf. on Parallel Processing, PP. 37-44, Aug 1998.
- [4] N. Neves and W. K. Fucks. « Adaptive Recovery for Mobile Environnements» ACM, vol 40, no 1, PP 68-74, January 1997.
- [5] G. Cao and M. Singhal. «*Mutable Checkpoints: A New Checkpointing Approach for Mobile Computing Systems* » Parallel and Distributed Systems, IEEE Transactions on Vol 12, Issue 2, Pages 157 – 172, Feb 2001
- [6] H. Higaki and M. Takizawa. « Checkpoint-Recovery Protocol for Reliable Mobile Systems ». The 17 th International Symposium on Reliable Distributed Systems, pp. 93-99. 1998.
- [7] C.M Lin et C.R Dow. « Efficient Checkpoint-Based Failure Recovery Technique in Mobile Computing Systems». Journal of Internat. Science and Engineering, vol 17, N°4, p549-573, Sept. 2001.
- [8] M. Aliouat, H. Mansouri, N. Badache, «Un Protocole efficace non-bloquant et optimal d'établissement coordonné de points de reprise dans les systèmes mobiles », conf. Intl. NOTERE, Marrakech 2007.
- [9] Sapna E. George, Ing-Ray et Chen Ying Jin. « Movement-Based Checkpointing and Logging for Recovery in Mobile Computing Systems » in 5th ACM International Workshop on Data Engineerin for Wireless and Mobile Access, p51-58, June 2006.
- [10] Bidyut Gupta and Shahram Rahimi. « A Novel Roll-Back Mechanism for Performance Enhancement of Asynchronous Checkpointing and Recovery ». Informatica Journal, vol 31, N°1, 2007.
- [11] S.K.Gupta, and R.K Chauhan and Parveen Kumar « A Non-intrusive Snapshot Collection Protocol for Distributed Mobile System » Intern Journal of Computer Science and Network Security, Vol 7 No.10, October 2007.
- [12] M. Aliouat, N. Labraoui, «Problématique du checkpointing dans les réseaux mobiles ad hoc » Conf. Intl de Productique, Sétif novembre 2007.

## **Node Disjoint Multipath Routing Protocol supporting Quality of Service in Wireless Sensor Networks**

Zahia BIDAI, Bouabdellah KECHAR, Hafid HAFFAF

Oran University, Faculty of Sciences, Department of Computer Science,  
BP 1524 El M'Naouar, ORAN, ALGERIA

{bidai\_zah, bkechar2000, haffaf\_hafid}@yahoo.fr

**Abstract:** Wireless sensor networks (WSN) are an important research area that has attached considerable attention for academic, industrial and institutional parts. The most works achieved until today, however, have concentrated in many cases on energy saving considerations and few ones on quality of service that characterize new applications such as reactive monitoring with reliability and delay constraints. In this paper we are interested by a time and data critical traffic. We propose an on demand Quality of Service (QoS) Multipath Routing Protocol called QMRP. We utilize multiple paths for reducing the routing overhead and providing reliability. The obtained results through simulation using NS2 simulator, by comparison with classical AODV, demonstrate the importance of multipath routing in terms of minimum routing overhead, reduction of lost packets for dynamic topology in WSN and the assurance of quality of service defined by the delay metric.

**Keywords:** Wireless sensor networks, Multipath routing, AODV, QoS.

### **1 Introduction**

The progress realized in the last few years in micro-electro-mechanical systems and wireless communication technology fields, allow to produce tiny components with a reasonable cost called sensor nodes. Each node is equipped with limited capacities in terms of processing, storage, communication and electric power supply. A wireless sensor network is regarded as a special type of ad hoc networks and consists of a great number of sensor nodes able to collect and transmit environmental data towards a base station, called Sink (collection point). In this paper, we consider a time and data critical class of traffic, for example, for a danger warning packet, where latency and reliability are important. In order to guarantee reliability and QoS requirement in terms of delay for the above traffic class, we propose an on demand multipath QoS extension of AODV [1]. Multipath routing alleviate single-path problems, such as high route discovery latency, frequent route discovery attempts and improve the performance of a WSN such as reliability, load/energy consumption balancing, routing overhead, quality of service and security [2]. Among these performance benefits from multipath routing, in this article, we are interested by two objectives: 1) reliability in terms of lost packets due to frequent topological changes, 2) reduction of routing overhead. To achieve this second goal, we have combined the path accumulation feature with delay constraint of QoS in order to optimize the flooding when discovering multiples paths. The concept of multipath routing consists in finding for each source node multiple choices of path enabling it to reach a particular destination (Sink). These multiple paths can be used either alternatively for reliable data delivery, i.e. at a given time only one path will be chosen, or simultaneously for load balancing by using several paths at the same time. This second case is out of scope of this paper. These paths can also be classified in three principal categories such as: link disjoint paths, not disjoint paths and node disjoint paths. In our proposal scheme, we are interested by the third category because the independence and resilience provided by the node disjoint paths. Depending on different and new applications, the traffic in WSN may be time critical (delay constraint) and/or data critical (reliability constraint). Therefore QoS routing is an important issue in WSN. Reliability is of great concern in WSN due to the fact that sensors are susceptible to failures. Many papers have proposed routing protocols to improve reliability of the networks which are based on multipath routing approach. Split Multipath Routing (SMR) proposed in [3] is an on-demand multipath source routing protocol based on DSR [4] and is used to construct maximally disjoint paths (paths with as few nodes or links in common as possible). AOMDV [5] is based on AODV [1] and can find link or node disjoint paths. AODV-Multipath (AODVM) [6] is also based on AODV, and can find only node disjoint paths. The Node-Disjoint Multipath Routing (NDMR) [7] protocol is also

based on AODV, it can efficiently discover multiple route paths between communicating nodes with minimal control overhead (low broadcast redundancy) and minimal routing latency. H-SPREAD [8] is a hybrid multipath scheme to improve security and reliability of data collection in WSN, this scheme is based on a distributed N-to-1 multipath discovery protocol which is able to find multiple node-disjoint paths from every sensor node to the Sink simultaneously in one route discovery process. On the other hand, a number of protocols have been proposed for QoS routing in WSN taking into account the delay and/or energy constraints. ADPROC [11] is an adaptive routing framework to provide QoS in WSN. It supports two different routes: a short rapid route for real-time data and energy efficient route for non-real-time. A QoS aware protocol for real-time traffic generated by a WSN consisting of image sensors is proposed in [12], the throughput for non-real-time data is maximized by adjusting the service rate for both real-time and non-real-time data at the sensor nodes. To the best of our knowledge, very few researches that consider reliability and delay QoS constraints in WSN exist. In MMSPEED [9] the authors propose a packet delivery mechanism for WSN to provide service differentiation and probabilistic QoS guarantees in timeliness and reliability domains. A multiconstrained QoS multipath routing in WSN is proposed in [10], the authors address both time and reliability constraints and formulate the problem in a deterministic linear programming. Our proposed protocol described in Section 3 adapts and extends AODV routing protocol for WSN and establishes multiple node disjoint paths by combining path accumulation strategy [7] and QoS delay constraint.

The outline for the remainder of the paper is as follows. In section 2, we describe our proposed multipath QoS routing protocol QMRP. Evaluation results of our proposal using NS2 simulator are presented and discussed in section 3, and finally section 4 concludes this paper.

## 2 Our Proposal

Our proposal scheme QMRP has two phases: Multiple Route Discovery phase and Route Maintenance phase, each of which is explained in the following sub-sections. QMRP uses three types of control messages and one type of data message DATA:

- Extended RREQ (Route REQuest) message contains eight fields (*type*: message type, *hop\_nbr*: number of hops from the source node to the current node, *broadcast\_id*: broadcast ID of the RREQ message, *ID\_dst*: destination node, *dst\_ns*: sequence number of destination, *ID\_src*: sequence number of source, *src\_ns*: source node, *TTL*: Time to Live value of RREQ message) followed by two extended fields (*list\_path*: contains the list of ID nodes of the path, *T\_delay*: maximum delay for a transmission from a source node or an intermediate node forwarding the RREQ packet to the sink node).
- Extended RREP (Route REPLY) message contains six fields (*type*, *Hop\_nbr*, *ID\_dst*, *dst\_ns*, *ID\_src*, *TTL*) explained above followed by the same extended field (*list\_path*).
- RERR (Route ERRor) message has the same AODV-RERR message format.
- DATA message has the same AODV-DATA message format.

Three types of table are used by each node in the WSN: 1) Neighbours table which contains the neighbours of the node, 2) Source\_broadcastID table used to record the ID of a source node in *ID\_src* field and the broadcast ID of the RREQ packet in the *Rreq\_ID* field and 3) Extended route table which contains forward and reverse routes entries. Each entry has three fields (*ID\_src*, *ID\_dst*, *dst\_ns*) described above and *Next\_hop*, *Next\_hop\_valid* and *Hop\_count* fields for each node disjoint path founded during the Multiple Route Discovery process (in our case the number of these paths is fixed to three at maximum), where *Next\_hop* is the next ID node of the route to the destination, *Next\_hop\_valid* is the flag that represents if the route is or not valid and *Hop\_count* the number of intermediate nodes from the source to the sink node on this route.

### 2.1 Multiple Route Discovery

In this sub-section, we explain the multiple route discovery process illustrated in figure 2.

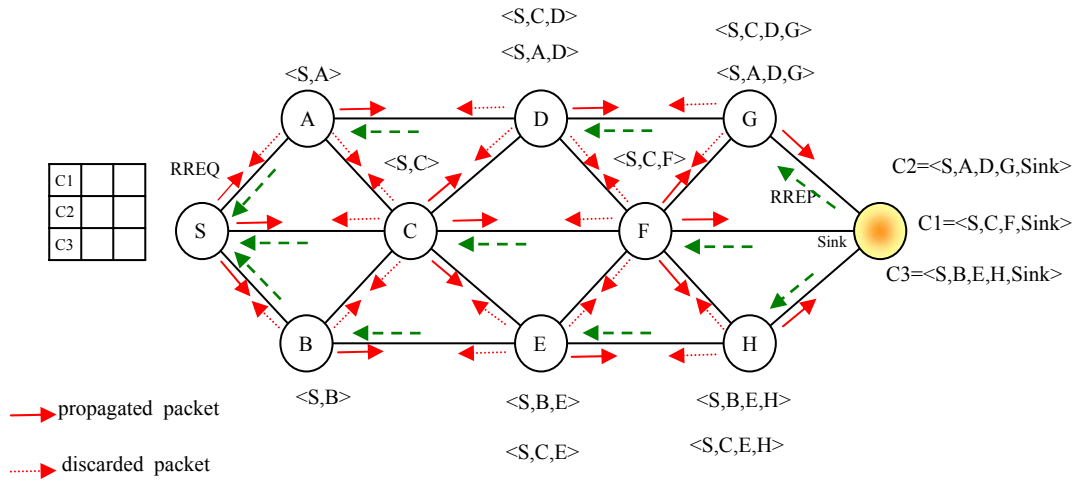


Fig. 2. Multiple route discovery process

**Source node task.** Initially, when a node source wants to send data to the Sink, and no route to this destination is available in its routing table, in this case, the source creates and broadcasts a RREQ packet (figure 1.a). The initial value of the  $T\_delay$  field represents the end-to-end delay requirement for a QoS path fixed by the target application.

**RREQ Packet processing.** We describe here the tasks of intermediate nodes and sink node when receiving RREQ message.

**Intermediate node task and Decreasing Broadcast Routing Overhead.** Overhead here is defined in terms of routing protocol control messages that consume both channel bandwidth as well as the battery power of nodes for communication/processing. So, decreasing the overhead is important in WSN where bandwidth and energy consumption constraints are limited. To reduce the broadcast routing overhead we have combined with a particular flooding based on the path accumulation approach [13] and the delay constraint. Moreover TTL value could be used to limit the initial flooding. When an intermediate node receives a RREQ packet, firstly, the node updates and checks the  $T\_delay$  value: it subtracts from the  $T\_delay$  indicated in the RREQ the time required by this node to process the RREQ, if  $T\_delay$  is not positive the node drops the RREQ packet. Otherwise, the node looks for the RREQ in the source\_broadcast table, if it is a new RREQ it records the  $hop\_nbr$  of the RREQ packet as the shortest number of hops in its reverse route entry of the route table (figure 1.c). If the node receives the RREQ duplicate again, it compares the RREQ  $hop\_nbr$  to the number of the shortest hops  $Hop\_count$  recorded in its reverse route entry of route table. If the number of hops is larger than the shortest number of hops, the node drops the RREQ packet. Otherwise, the node appends its own ID to the  $list\_path$  of the RREQ packet and broadcasts the RREQ packet to its neighbouring nodes. In figure 2, from source node S to node C there are six paths:  $\langle S, C \rangle$  (1 hop),  $\langle S, A, C \rangle$  (2 hops),  $\langle S, B, C \rangle$  (2 hops),  $\langle S, A, D, C \rangle$  (3 hops),  $\langle S, B, E, C \rangle$  (3 hops),  $\langle S, C, F, C \rangle$  (loop path with 3 hops), the node C records only 1 as the shortest number of hops and discards the other five RREQ packets. Therefore, only part of RREQ packets arrive at sink node when comparing with the flooding manner. This combined approach reduces overhead and avoids forming loop paths.

**Destination node task and selecting node-disjoint paths.** Not all paths that reach the sink node are node disjoint. The sink must select only the node disjoint paths among these: When the destination node (sink) receives the first RREQ packet, it records the  $list\_path$  which contains the list of node IDs for the entire route path in its cache and sends a RREP packet (figure 1.b) that includes the  $list\_path$  towards the source node. For the duplicate RREQ packets received by the destination and which are not all node-disjoint paths, the destination compares the  $list\_path$  of the received RREQ to all the existing node-disjoint paths in its cache. If there is not a common node (except source and destination) between the  $list\_path$  and any node-disjoint path recorded in the destination's cache memory, the path is selected as a node-disjoint path and is recorded in the destination's cache memory. Otherwise the RREQ packet is dropped.



**RREP packet processing.** We describe here the tasks of intermediate nodes and source node when receiving RREP message.

**Intermediate node task.** When an intermediate node receives a RREP packet, it updates its routing table from the *list\_path* of the RREP packet: it records the node ID of the destination and the node ID of the neighbour from which the RREP arrived in the forward route entry of the route table. The node ID of the source and the node ID of the next hop to the source are recorded in the reverse route entry of the route table. Finally the intermediate node forwards the RREP towards the source node along the reverse route path.

**Source node task.** When the first RREP packet arrives at the source node, it records the next hop to the destination node in the forward route entry and can use this new path to send data packets. For the other RREP packets which arrive later to the source node, it records them in the route table as alternate path used when a broken link is detected. The number of node disjoint paths recorded in the route table of the source node is limited to three paths due to the limited storage capacity of a node in WSN. To decrease packet overhead, the *list\_path* is utilized only during multiple route discovery, not during data transfer phase.

## 2.2 Route Maintenance

When a node cannot receive HELLO messages from a neighbor node, the node detects link rupture due to the mobility of nodes and sends a RERR message to the source node along the reverse route path. When an intermediate node receives a RERR packet, it marks its route to the destination node invalid and then propagates the RERR to its precursor node along the reverse route path. After receiving the RERR, the source node invalidates the route path to the destination node and changes a primary route to a valid alternate node-disjoint path from the routing table to continue to forward data packets or initiates a new multiple route discovery process when no alternate route are available.

## 3 Performance Evaluation

The effectiveness of the multipath routing approach is validated through simulation. We consider three different routing protocols for operational comparison: 1) AODV routing protocol, 2) QMRP Multipath Routing Protocol providing QoS in terms of delay (our proposed routing protocol) and 3) MRP the multipath version of QMRP without delay constraint. This section describes the simulation environment, performance metrics and preliminary experimental results.

### 3.1 Simulation Environment

We evaluated our proposal using NS-2.31 simulator [11]. A total of 10 simulations are conducted with the number of sensor nodes ranging from 30-120. A simulated field is 1500[m]\*1500[m] inside which nodes can move with low velocity. In this paper, we consider only one sensor node is actively sending data to the sink. The rest of parameters are summarized in table 1.

**Table1.** System parameters

Channel Type	Wireless channel
Propagation model	Two ray ground
MAC layer	CSMA/CA IEEE 802.11
Transmission Rate	32 kb/s
Queue management politic	DropTail
Maximum of queue	30 packets
Radio Range	250 m
Routing Protocol	AODV, MRP and QMRP
Traffic Model	UDP/CBR

### 3.2 Performance Metrics

Performance metrics are as follows:

- Number of routing messages: is the number of messages used by the routing protocol to discover the paths.
- Number of lost packets: is the number of data packets which did not reach their destinations.

### 3.3 Performance Results

**Number of routing messages.** In order to show the performance of our approach in terms of reduction of the control packets number in the network, which influences directly network lifetime and the network reliability, we analyzed the number of routing messages according to three parameters: without link rupture, with only one link rupture and with  $N$  link ruptures.

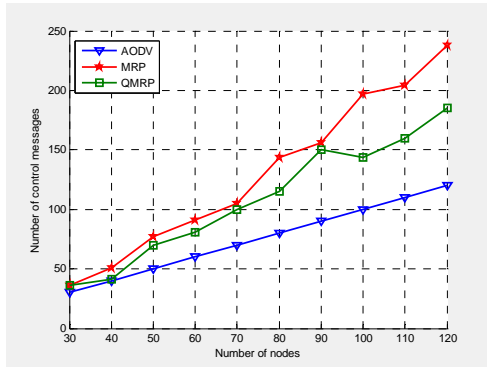


Fig. 3. Number of control messages without link rupture.

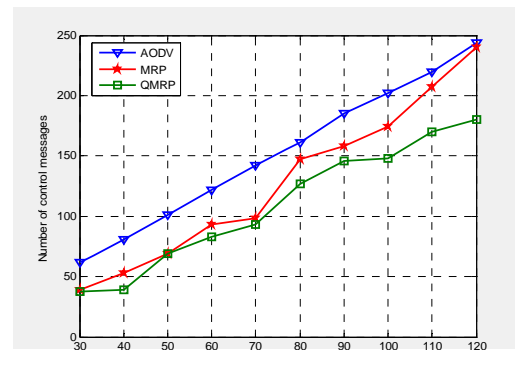


Fig. 4. Number of control messages with one link rupture.

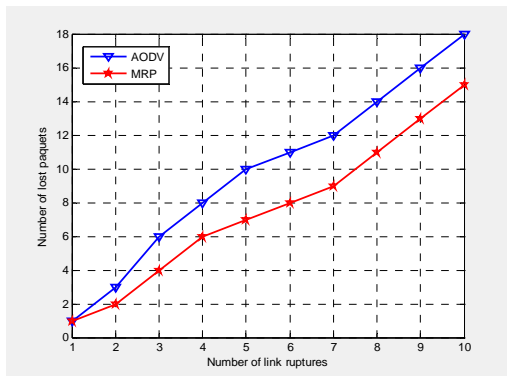


Fig 5. Number of control messages with  $n$  link ruptures.

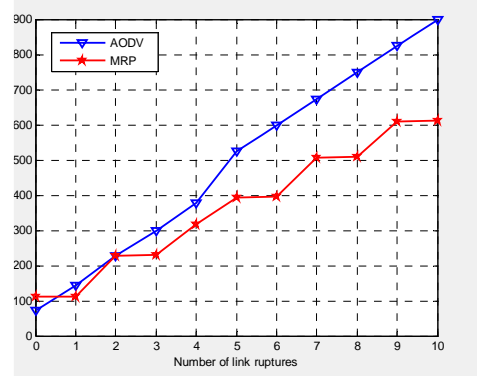


Fig. 6. Number of lost packets with  $n$  link ruptures.

Figure 3 shows that the control messages number of MRP protocol increases more quickly than AODV. That can be explained by the great diffusion of RREQ packets in MRP to increase the possibility of having node disjoint paths which are useless in this case (without link rupture). But this number decreases in QMRP which discarded every RREQ packets that don't respect the delay constraint.

Figures 4 and 5 show that MRP has low number of control messages than AODV but more than QMRP. MRP can find multiple node-disjoint paths in a route discovery process, so the protocol decreases the number of path rediscovery processes. On the contrary, since AODV encounter (one or more) link ruptures with the increase of mobility, it has to trigger more new route discovery process which cause more routing control messages to be sent to the whole network. These results show that multipath approach routing is important and robust faces link failures. Moreover, QMRP reduces even more the number of control packets. The reason is that RREQ messages which don't respect the delay constraint

are discarded. This reduction of control messages number can, on one hand, guarantee enough available bandwidth which means that the delay takes by a packet to reach the sink node decrease. On the other hand, allow the conservation of node energy and thus the increase of node lifetime.

**Number of lost packets.** Figure 6 gives variation of the number of lost packets as a function of the number of link ruptures. As the number of link failures increases (0 to 10), the number of lost packets increases. MRP has low lost packets than AODV, the reason is that MRP has multiple node disjoint paths. When an active path is broken due to link rupture, the source node that generate data will receive a RERR message, so it invalidates the broken path and choose quickly another valid node disjoint path from its route table in order to avoid the loss of packets and continue to transmit data between source and destination .

## 4 Conclusion

In this paper an on-demand node disjoint multipath routing protocol supporting quality of service in wireless sensor networks QRMP has been proposed. QRMP provides reliability and QoS under delay constraint. A combined approach based on path accumulation and delay constraint is used to reduce the control messages overhead dramatically and determine multiple node-disjoint routing paths during a single multiple route discovery. The preliminary results of simulation indicate that the proposed protocol QRMP and MRP outperform AODV in terms of lost packets number and control messages overhead. This shows clearly that a multipath approach provides robustness to frequent topological changes which characterizes WSN and is very interesting and promising routing technique, especially if QoS is integrated. The encouraging preliminary results obtained in this work provide the first step research. Other QoS metrics including end to end delay, residual energy consumption, interferences and congestion control can be added to examine their performance benefits from multipath routing approach used in the QRMP routing protocol.

## References

1. Perkins, C., Belding-Royer, E., Das, S.: Ad hoc On-Demand Distance Vector (AODV) Routing. Network Working Group, July 2003, <ftp://ftp.nordu.net/rfc/rfc3561.txt>
2. Lou, W., Liu, Y., Zhang, Y.: Performance Optimization using Multipath Routing in Mobile Ad Hoc and Wireless Sensor Networks. Combinatorial Optimization In Communication Networks, Kluwer Academic Publishers, 2005.
3. Lee, S.-J., Gerla, M.: Split multipath routing with maximally disjoint paths in ad hoc networks. IEEE International Conference on Communications, Volume 10, June 2001.
4. Johnson, D.B., Maltz, D.A.: Dynamic source routing in ad hoc wireless networks. Mobile Computing, 1996, pp. 153-181.
5. Marina, M.K., Das, S.R.: On-demand multipath distance vector routing in ad hoc networks. Proceedings of the International Conference for Network Protocols (ICNP), November 2001.
6. Ye, Z., Krishnamurthy, S.V., Tripathi, S.K.: A framework for reliable routing in mobile ad hoc networks. IEEE INFOCOM, 2003.
7. Li, X.: Multipath Routing and QoS Provisioning in Mobile Ad hoc Networks. Thesis, Department of Electronic Engineering Queen Mary, University of London, April 2006.
8. Lou, W., Kwon, Y.: H-SPREAD: A Hybrid Multipath Scheme for Secure and Reliable Data Collection in Wireless Sensor Networks. Appeared in Vehicular Technology, IEEE Transactions on, July 2006, volume 55, issue 4, on pages 1320-1330.
9. Felemban, E., Lee, C.-G., Ekici, E., Boder, R., Vural, S.: Probabilistic QoS guarantee in reliability and timeliness domains in wireless sensor networks. Proceedings of 24th Annual Joint Conference of the IEEE Computer and Communications (INFOCOM), pages 2646-2657, March, 2005.
10. Huang, X., Fang, Y.: Multiconstrained QoS multipath routing in wireless sensor. Wireless network, Springer Science + Business Media, LLC January 2007.
11. The ns manual. <http://www.isi.edu/nsnam/ns/tutorial/>.
12. Gundappachickenahalli, K., Ali, H.: ADPROC: An Adaptive Routing Framework To Provide QoS In Wireless Sensor Networks. Proceedings of the 24<sup>th</sup> IASTED International Multi-Conference, Parallel and Distributed computing and networks, February 14-16, 2006, Innsbruck, Austria.
13. Akkaya, K., Younis, M.: An Energy-Aware QoS Routing Protocol for Wireless Sensor Network.. in: Proc. Of Intl. Conf. on Distributed Computing Systems Workshops (ICSDSW), Washington, DC, 2003.

*E-learning II*

---



## Modélisation de l'apprenant dans les systèmes hypermédias adaptatifs

Amel Behaz \* — Mahieddine Djoudi \*\*

\* Faculté des Sciences, Université de Batna (05000) Algérie

\*\* Laboratoire SIC et Equipe IRMA

UFR Sciences SP2MI, Université de Poitiers

Teleport 2, boulevard Marie et pierre Curie,

BP 30179, 86962 Futuroscope Chasseneuil Cedex – France

mahieddine.djoudi@univ-poitiers.fr ; amelbehaz@hotmail.com

**RÉSUMÉ.** Tout processus d'adaptation repose sur un modèle de l'utilisateur, une représentation de ses caractéristiques dont le système prend en compte. Les systèmes hypermédia adaptatifs sont basés sur le profil d'apprenant, les données de plus en plus nombreuses et volumineuses doivent s'adapter aux apprenants (connaissance générales, buts, niveau de connaissances et préférences) et à leurs contextes d'utilisation. Dans ce papier, Nous étudions en détail les modèles qui nous ont paru les plus intéressants. Enfin, nous proposons notre modèle apprenant.

**MOTS-CLÉS :** Système hypermédia adaptatif, Modélisation, Profile apprenant

### 1. Introduction

Le but de la modélisation de l'apprenant est de donner une description aussi complète et fidèle que possible de tous les aspects relatifs aux comportements de cet utilisateur et du système qu'il utilise. La modélisation est une étape préalable pour l'élaboration des méta-données. Cette modélisation permet une adaptation qui améliorera le fonctionnement et la convivialité du système, ainsi ce dernier peut présenter les informations les plus intéressantes et aider l'utilisateur dans sa tâche d'apprentissage.

Nous présentons nos travaux concernant la modélisation des apprenants dans les hypermédias adaptatifs. Ces travaux ont pour but d'offrir un cadre général pour la conception de modèles de l'utilisateur spécifiques à un domaine d'application.

Dans cet article, nous présentons d'abord un état de l'art des différentes façons d'envisager une modélisation des utilisateurs dans un système hypermédia, les propriétés et techniques d'acquisition. La dernière section est consacrée à la description du modèle apprenant adopté dans cette étude.

### 2. Classification

Dans les systèmes hypermédias adaptatifs, le modèle utilisateur reste assez générale, on retrouve le plus souvent 2 types de données, les connaissances que possède l'utilisateur sur le domaine qui sont fondamentaux et les connaissances générales concernant l'utilisateur lui-même qui ne sont pas en rapport avec le domaine. La classification des modèles utilisateurs est basée sur 3 dimensions:

- Un Modèle canonique : consiste à utiliser une représentation appelée stéréotype des besoins, des caractéristiques, du fonctionnement cognitif, du mode de raisonnement, etc. de l'ensemble des utilisateurs (Koch, 2000).

- Modèles basés sur des caractéristiques persistantes ou ponctuelles qui permettent de distinguer entre les informations valides à long terme et celles à court terme (Koch, 2000).

### **3. Description des caractéristiques d'un apprenant**

Tout processus d'adaptation, repose sur un modèle de l'apprenant et une représentation de ses caractéristiques dans le système. Nous décrivons chaque caractéristique :

#### **3.1. Les Connaissances de l'apprenant**

C'est la donnée la plus importante dans tout système adaptatif. Il s'agit de sauvegarder ce qu'un apprenant connaît ou ne connaît pas par rapport aux concepts du domaine. Cette donnée peut être obtenue explicitement suite à un test par exemple ou bien implicitement par application du système des règles d'inférences. Les connaissances peuvent varier, pour le même apprenant, d'un moment à un autre. En effet, à chaque utilisation du système l'apprenant peut acquérir de nouvelles connaissances et augmenter son niveau ou au contraire le diminuer. Il est donc nécessaire de disposer d'outils d'évaluations robustes afin de confirmer les capacités effectives.

#### **3.2. Les préférences**

Contrairement à la donnée précédente cette caractéristique ne peut pas être déduite par le système. Chaque apprenant possède ses propres préférences, il doit les communiquer au système. Les préférences se divisent en deux catégories les préférences d'affichage taille des caractères, problèmes de couleurs, de contraste, etc., et les préférences de présentation. Ainsi, il a possibilité par exemple de définir un classement sur les types physiques de média pour la présentation des contenus (texte, image, vidéo,...).

#### **3.3. Les objectifs et plans**

(Laroussi 2001) Définit un objectif comme étant l'état que l'apprenant souhaite atteindre et les plans décrivent les étapes pour y arriver. Ces informations peuvent être données explicitement par l'apprenant ou bien déduites par le système grâce à certaines règles d'organisation de la structure narrative d'un document.

#### **3.4. Historiques**

Les historiques regroupent les actions que l'utilisateur a effectuées dans le système, des concepts abordés ou des documents parcourus. Cette caractéristique n'est pas toujours utilisée. En revanche, elle peut être nécessaire pour représenter plusieurs sortes d'historiques

### **4. Les modèles de représentations des connaissances**

Différentes techniques sont utilisées pour représenter les connaissances des apprenants dans un système hypermédia.

**4.1. Modèles de recouvrements :** (overlay models) Dans ce type de modèle il s'agit de modéliser la connaissance d'un individu et non celle d'un groupe. Les connaissances sont représentées par un sous ensemble de connaissances. Ces dernières sont généralement exprimées par un modèle et préalablement établi par un expert du domaine (définir ce que l'on attend de l'utilisateur, les connaissances qu'on souhaite le voir acquérir). Cet ensemble prend la forme de paires concept-valeur ou concept représente la connaissance, valeur le degré de maîtrise que l'apprenant possède par rapport à ce concept. Valeur peut être (booléenne, nominale, une probabilité....) c'est un modèle puissant et flexible. On peut évoluer facilement le niveau de connaissance de l'apprenant pour chaque concept.

**4.2. Modèles stéréotypes :** Permettent la modélisation de la connaissance dans un groupe. L'apprenant est classé sous un stéréotype. Ces modèles sont fondés sur des paires items-valeurs qui définissent différents stéréotypes tel que novice, avancé, expert. Dans ce cas l'apprenant est associé à une de ces catégories, hérite de ses propriétés et dispose des adaptations élaborées par le stéréotype en question (Iksal, 2001).

On remarque que dans les modèles stéréotypes le manque d'individualisation de l'adaptation. En effet, le modèle stéréotype propose le niveau minimal qu'une classe d'apprenants dispose sur un concept donné. Par contre le modèle de recouvrement représente exactement le niveau de connaissances de chaque individu dans le système. Il serait intéressant d'élaborer un modèle mixte des deux approches. Initialiser le modèle de l'apprenant à l'aide d'un stéréotype, puis l'affiner en utilisant un modèle de recouvrement (le réévaluer de nouveau en fonction de ses actions et réponses utilisant une méthode probabiliste)

Nous avons vu que l'adaptation nécessite des connaissances sur les utilisateurs. Ces connaissances sont construites selon des modèles stéréotypes (classe d'individus) ou des modèles de recouvrement (individu). Par ailleurs, il existe deux manières d'obtenir de l'information sur les utilisateurs : l'acquisition explicite selon laquelle des interactions avec le système permettent de créer ou compléter le modèle. L'acquisition implicite selon laquelle le système infère l'information suivant le comportement de l'utilisateur et des règles d'inférences. Le tableau 1 récapitule tous ces points.

Modèle Apprenant		
Caractéristiques	Modèle de Représentation	Acquisition
Connaissance	Stéréotype/Recouvrement	Implicite/ Explicite
Objectifs et plans	Stéréotype/Recouvrement	Implicite/ Explicite
Styles d'apprentissage	Stéréotype	Implicite
Préférences	Stéréotype	Explicite
Historiques	Recouvrement	Implicite

**Tableau 1.** Récapitulation des modèles de représentation

## 5. Un modèle apprenant

Dans cette section, nous allons détailler le modèle apprenant que nous proposons pour le e-learning.

### 5. 1. Caractéristiques du modèle

– La classe **IdentiUser** est chargée de représenter les informations concernant un utilisateur particulier. Elle est composée d'attributs prédéfinis indispensables et communs à tous les utilisateurs : **nom, prénom, identifiant, mot de passe, âge, organisme...**

– La classe **Préférence** chargée de représenter les préférences d'un apprenant. En spécifiant les types de média préférés. Identifiée par **nom** et une **valeur**

– La classe **Capacité** (connaissance) chargée de représenter ou de donner un niveau de connaissance d'un apprenant sur un concept donné les valeurs possibles sont : **très bas, bas, moyen, bon, excellent** cette échelle nous permet d'avoir plus de précision sans qu'elle soit trop profonde, nécessaire à l'adaptation.

– La classe **Historique** est chargée de garder trace de l'état des historiques d'un apprenant. Mémorisation de la navigation et des ressources lues dans les documents. Cette représentation permet de donner la date du parcours d'un document par exemple ou encore les chemins de parcours dans l'ordre de navigation (méthodes utilisées : **DocSuiv ()** et **DocPréc ()**).

– La classe **Paramètres** chargée de contenir les noms des stéréotypes utilisés ainsi que la liste de leurs valeurs possibles (**vitesse d'apprentissage** : *lent* ou *rapide*) (**stratégie** : *ascendante* ou *descendante*)

La plupart des modèles utilisent une seule classe pour représenter tous ces attributs. Nous avons choisi de représenter des classes différentes car nous croyons qu'elles engendrent des sémantiques différentes (Figure 1). Tous ces attributs sont souvent utilisés dans les hypermédias adaptatifs à but pédagogique, ils sont indispensables dans la conception. Contrairement à certains modèles, la classe historique nous paraît indispensable : car elle va permettre à un apprenant de parcourir à nouveau des documents déjà parcourus, composés des mêmes ressources atomiques que lors du premier parcours. Il serait facile de retrouver les documents parcourus qui ont permis à un apprenant d'acquérir un concept donné (noeuds visités). Les classes Document et Concept sont détaillés dans nos travaux (Behaz, 2005) et (Behaz, 2007) le modèle du domaine. La classe Gestionnaire des Interactions est chargée de servir d'interface avec les autres composants du système d'hypermédia adaptatif. Elle permet de poser des questions à un utilisateur afin de collecter les données nécessaires et d'apporter des réponses aux composants externes (modèle du domaine, modèle de l'adaptation).

## 5.2. Eléments de Comparaison avec d'autres modèles Utilisateurs

Dans cette partie nous allons essayer de comparer notre modèle avec certains modèles qui existent déjà. Certes, nous avons construit ce modèle en nous inspirant des modèles existants. Néanmoins, il propose un cadre général pour concevoir des modèles de l'utilisateur spécifiques pour n'importe quel domaine d'application.

Les modèles standards type IMS Learner (IMS, 2001) ou PAPI (PAPI, 2002) organisent un certain nombre de données sur l'utilisateur selon différentes structures. Ces données constituent un exemple complet et normalisé d'informations que l'on peut vouloir stocker dans un système d'hypermédia adaptatif. Dans les deux modèles, la totalité des champs sont optionnels. Notre modèle est bien compatible avec ces standards car il réutilise un sous ensemble d'attributs sélectionnés.

Les modèles Munich Model (Koch, 2001), INT (Duitama, 2005) et AHAM (De Bra, 2004) qui sont aussi des modèles généraux. Seulement, ils présentent quelques différences avec notre modélisation.

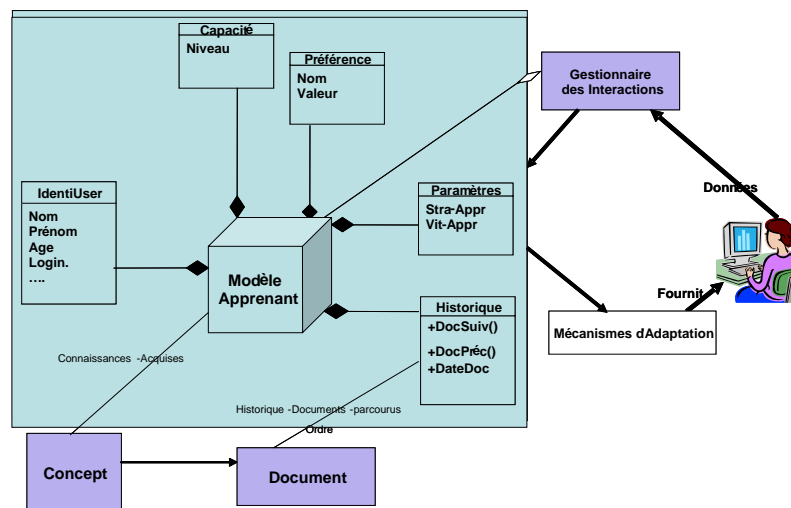


Figure 1. Modèle Apprenant élaboré

### 5.2.1 Le Munich Model

Le modèle Munich (De Bra, 2002) est composé d'une classe gestionnaire d'utilisateurs, qui est composée d'un ou de plusieurs utilisateurs. Un utilisateur possède des attributs nom d'utilisateur et email.



Il contient également d'autres attributs, potentiellement quelconques. Les attributs sont divisés en deux, ceux qui dépendent du domaine, et ceux qui n'en dépendent pas. Un attribut est constitué d'un nom et d'une valeur. L'identifiant est unique pour un utilisateur dans toute l'application. Les attributs dépendant du domaine sont liés aux composants du domaine. Ce modèle a une seule catégorie d'attributs indépendants du domaine. Dans notre cas nous en avons plusieurs, afin de pouvoir raisonner différemment sur différentes sortes d'attributs. Ce modèle ne permet pas de conserver la trace des documents parcourus : il n'y a pas d'historique, dans notre cas si.

### 5.2.2. Modèle INT

Dans sa récente thèse (Duitama, 2005) décrit le modèle de l'apprenant de la façon suivante. Le modèle de l'utilisateur est un triplet :  $UM = \langle learner, preference, knowledge \rangle$  learner est l'identifiant de l'utilisateur. Les préférences de l'utilisateur correspondent à sa langue ou à sa présentation préférée. Chaque élément de l'ensemble des préférences est représenté par une paire  $\langle attribute, value \rangle$ . Les connaissances de l'utilisateur sont représentées par des triplets de la forme :  $\langle domainConcept, role, educationalState \rangle$ . Là encore, les connaissances sont séparées des attributs indépendants du domaine. Les attributs indépendants du domaine sont représentés par une paire nom/valeur (Jacquot, 2006). Les attributs de connaissance possèdent une valeur (*educational-state*) et un rôle, correspondant au type de la connaissance acquise (connaissance pratique, théorique, etc).

### 5.2.3. Modèle AHAM

Le modèle de l'utilisateur de AHAM est relativement peu formalisé (De Bra, 1999). Il reste très général. Dans AHAM, le modèle de l'utilisateur est constitué d'attributs, qui correspondent le plus souvent à la connaissance de concepts. Pour chaque entité, un certain nombre de paires attribut-valeur sont stockées, comme, par exemple, le fait qu'un concept ait été lu ou le niveau de connaissance d'un concept. Les données relatives aux utilisateurs sont stockées dans une table. Cette table contient l'état des connaissances que l'utilisateur a du domaine, ainsi que ses préférences si nécessaire. La façon dont les caractéristiques des utilisateurs sont mises en relation avec les autres éléments du modèle est implicite. Il n'y a pas d'historique. Il n'y a pas d'attributs prédéfinis.

## 6. Conclusion

Il est possible d'utiliser les attributs que l'on souhaite, sans contrainte. Il suffit de choisir d'utiliser les structures et le vocabulaire des standards IMS et PAPI pour former les paires attribut-valeur nécessaires à la constitution du modèle. On note également qu'il n'y a pas toujours un fond commun d'attributs absolument indispensables pour tous les modèles de l'utilisateur, ou alors ces attributs sont très peu nombreux. Cependant, il apparaît que les notions de but, d'historique, d'identifiant, etc. sont très utiles dans les HA. Certains aspects des modèles de l'utilisateur ne sont pas toujours pris en compte dans les modèles formels, notamment la façon dont les stéréotypes sont calculés.

Pour concevoir un modèle d'utilisateur pour un domaine particulier, il semble donc utile d'avoir à disposition un certain nombre d'attributs prédéfinis, dont certains peuvent être optionnels, et de donner un cadre pour faciliter la création d'attributs non-prédéfinis. Il est difficile de fournir à l'utilisateur final l'information dont il a besoin à l'instant précis où il en a besoin. L'accès à une information pertinente, adaptée aux besoins et au contexte de l'utilisateur est un vrai challenge.

## 7. Bibliographie

- Behaz A., Djoudi D., Génération dynamique de documents hypermédias adaptatifs dans un environnement numérique de travail, *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées (ARIMA)*, ISSN 1638-5713, Vol. 3, November 2005,
- Behaz A., Djoudi D., « Création dynamique de documents hypermédias adaptatifs », *10<sup>ème</sup> Colloque International sur le Document Electronique, CIDE'10*, 2 - 4 juillet 2007
- De Bra P., Houben G., Wu H., AHAM : A dexter-based reference model for adaptive hypermedia. In UK Conference on Hypertext, p 147–156, 1999.
- De Bra P., Brusilovsky P., and Conejo R., Adaptive Hypermedia and Adaptive Web-Based Systems, Second International Conference, AH 2002, Malaga, Spain, May 29- 31, 2002.
- De Bra P., Nejdl W., « Adaptive Hypermedia and Adaptive Web-Based Systems », *Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23- 26, 2004*.
- Duitama F., Defude B., Bouzeghoub A., and Carpentier C., « A framework for the generation of adaptive courses based on semantic metadata », *Multimedia Tools and Applications*, 25-3 :377–390, 2005.
- Iksal S., Garlatti S., Ganier F., and Tanguy P. « Semantic composition of special reports on the web. a cognitive approach ». In Balpe J.-P., Leleu-Merviel S., Saleh I., and Laubin J.-M., editors, *Actes de H2PTM'01*, pages 363–378. Hermès Science Publications, Hypertextes, hypermédias : nouvelles écritures, nouveaux langages, 2001.
- IMS <http://www.imsglobal.org/profiles/index.html>. Ims learner information package specification.
- Jacquot C., Modélisation logique et générique des systèmes d'hypermédias adaptatifs, Thèse de doctorat, Université Paris XI Faculté des sciences d'Orsay, 2006.
- Koch N., Software Engineering for Adaptive Hypermedia Systems Reference Model, Modelling Techniques and development process. These de doctorat, Factulé de mathématiques et Informatique, Université Ludwig-Maximilians, Munich, 2000.
- Koch N., Wirsing M., The Munich reference model for adaptive hypermedia applications. pages 213–222, 2001.
- Laroussi M., Conception et realization d'un système dydactique hypermedia adaptatif : CAMELEON. These de doctorat, Université Manouba, Tunisie, 2001
- Papi Papi learner, draft 8 specification. <http://edutool.com/papi/>.

## Vers une adaptabilité du scénario d'apprentissage à base d'indicateurs de parcours

Samia Ait Adda<sup>1</sup>, Nabila Bousbia<sup>1</sup>, Amar Balla<sup>1</sup>

<sup>1</sup> Institut National d'Informatique, INI

BP 68M, Oued-Smar, CP 16309, Alger, Algérie

[{s\\_ait\\_adda,n\\_bousbia,a\\_balla}@ini.dz](mailto:s_ait_adda,n_bousbia,a_balla@ini.dz)

**Résumé.** Depuis quelques années, les environnements informatiques pour l'apprentissage humain ont rapidement évolué grâce à l'évolution des technologies de l'information et de la communication. Toutefois, avec le nombre croissant d'utilisateurs de ces systèmes et la différenciation de l'espace et du temps entre les apprenants et les enseignants, le suivi et l'adaptation des contenus deviennent de plus en plus difficiles. Pour faire face à ce problème, plusieurs travaux de recherche ont été établis dans l'objectif d'adapter le contenu par l'interprétation des comportements des apprenants. Dans cet article, nous nous situons au niveau de l'adaptation des scénarios d'apprentissage. Nous essayons d'offrir à l'enseignant un moyen pour comparer le scénario prescrit avec celui réalisé. Notre approche est basée sur la collecte d'informations durant la session d'apprentissage, sur le parcours de l'apprenant. Ces informations seront structurées en indicateurs donnant à la fois une interprétation sur le comportement de l'apprenant et ses résultats permettant à l'enseignant de revoir le contenu pour l'adapter aux apprenants.

**Mot-clés :** scénario pédagogique, adaptation, trace, analyse de l'utilisation, indicateur.

### 1 Introduction

L'adaptation des environnements informatiques pour l'apprentissage humain (EIAH) est l'un des principaux verrous de leur évolution. L'enjeu est de comprendre le comportement et l'attitude de l'apprenant, ou d'un groupe d'apprenants, qui utilise un EIAH. C'est aussi, de pouvoir fournir à l'enseignant concepteur ou tuteur une information précise et adéquate pour ses besoins propres sur l'évolution individuelle et collective des apprenants afin d'adapter les scénarios d'apprentissage. La résolution de ce problème d'adaptation est essentiellement dépendante de la capacité à produire des traces pertinentes et exploitables sur l'activité d'un apprenant qui interagit avec un système d'apprentissage.

Dans cet article nous proposons une démarche d'aide à l'adaptation des scénarios pédagogiques en se basant sur des indicateurs de parcours de l'apprenant qui sont construits à partir des traces collectées lors de son apprentissage. Les valeurs d'indicateurs, ainsi calculer au niveau scénario pédagogique, participent à la régulation de la situation d'apprentissage et la révision du cours de la part de l'enseignant (concepteur ou bien tuteur) qui sera de ce fait l'acteur adaptateur.

Nous présentons d'abord un volet sur les traces, leur traitement et exploitation. Nous arborons ensuite le besoin de construire des indicateurs afin d'aider les différents acteurs à réaliser l'adaptation. A ce stade nous présentons notre solution.

### 2 L'adaptabilité du scénario d'apprentissage

Une des évolutions des EIAH est la possibilité de scénariser l'activité pédagogique de l'apprenant [3]. Ce progrès doit permettre à l'enseignant de définir des situations d'apprentissage répondants à ses objectifs pédagogiques et adaptables aux apprenants. Cependant, cette notion de scénario prête à la confusion dans le contexte éducatif : de la simple préparation du cours à la description détaillée du plan d'une séquence pédagogique d'activités, en passant par l'énoncé d'un travail préconisé aux apprenants.

Cependant, les scénarios ne peuvent être définitivement ou totalement définis lors de la conception initiale. Un certain nombre de facteurs peuvent conduire à laisser des imprécisions, ou à modifier certaines caractéristiques durant leur exploitation. Partant de ces constats, nous situons notre

problématique dans la mise à disposition d'environnements permettant aux enseignants la conception, le suivi ainsi que la modification de scénarios, afin d'assurer l'adaptation et la régulation des situations d'apprentissage.

Cette adaptation peut être effectuée en fonction des observables de la situation que nous appelons trace d'apprentissage. Ces observables permettent à l'enseignant d'avoir un retour d'information sur les activités de l'apprenant : positionner son avancement dans le scénario, voir sa progression et toucher ses manques et besoins.

Pour mieux comprendre où situer l'observation, le volet suivant donne une vue d'ensemble de ce concept et son rôle pour l'adaptation et la régulation.

### 3 Observation des comportements

Dans une activité d'apprentissage, l'objectif principal de l'observation est d'être un outil de support et de gestion chez l'apprenant, qu'il soit en activité individuelle ou collective, ou chez l'enseignant qu'il soit concepteur du cours ou bien tuteur. La trace collectée durant la session d'apprentissage peut être organisée dans un modèle afin d'être diagnostiquée, analysée puis exploitée.

Néanmoins, il ne faut pas limiter le sens de trace en une séquence d'actions, mais d'une manière générale, en une donnée qui fournit de l'information sur les sessions d'apprentissage. Une trace est une trace d'activité, d'utilisation ou d'interaction. Il ne lui est pas associé d'interprétation sur la situation d'apprentissage que lorsqu'elle participe à l'extraction et la construction des informations dites « de plus haut niveau », que nous appelons indicateur d'apprentissage. Après avoir été structurées et traitées, les traces primaires et les indicateurs peuvent enfin être exploités.

En fonction de son rôle, chaque acteur de la situation d'apprentissage pourra : analyser et adapter les caractéristiques identifiables (activité d'observation) ; établir un diagnostic sur les résultats et les progrès observés (activité d'évaluation) ; agir sur la situation d'apprentissage (activité de régulation). A ces fins, nous allons proposer notre solution dans le paragraphe qui suit.

### 4. Détection des comportements à base d'indicateurs d'apprentissage

Nous avons vu dans les paragraphes précédents l'utilité des traces et des indicateurs pour l'interprétation du comportement de l'apprenant afin d'adapter la situation d'apprentissage particulièrement le scénario pédagogique. Nous proposons de développer dans cette partie des indicateurs de parcours, mais avant cela, nous allons d'abord présenter le schéma général de notre solution, par la suite le modèle adopté pour la construction d'indicateurs, enfin nous détaillons notre solution.

#### 4.1 Schéma Général de l'Outil de Gestion de Trace

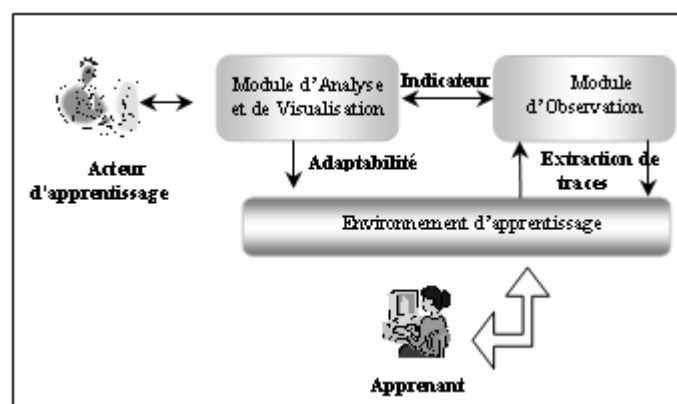


Fig.2. Schéma général de la solution

Durant l'interaction de l'apprenant avec l'environnement d'apprentissage, il produit des traces qui seront extraites via un outil de collecte installé à son niveau. Cet outil récupère les URL des pages visualisées, le temps de consultation de chaque page ainsi que les actions effectuées sur les pages (l'utilisation des ascenseurs, la sélection de texte suivi d'une action du type copier, etc.). Ces informations seront transmises au module d'analyse pour calculer les indicateurs demandés par l'enseignant. En effet, avant la phase d'observation, l'enseignant (concepteur/tuteur) sélectionne les indicateurs qu'il veut visualiser sur le parcours de l'apprenant, grâce à une interface du module d'analyse et de visualisation. Chaque indicateur étant déjà modélisé avec un format propre à l'objectif pour lequel il est conçu. Ce modèle sera présenté dans le paragraphe qui suit définissant la structure de l'indicateur et ses dépendances pour le calcul de sa valeur.

Après la phase d'analyse, l'indicateur aura une valeur qui sera avisée directement à l'apprenant et/ou l'enseignant, via une l'interface de visualisation. L'utilisateur pourra ainsi l'interpréter pour l'utiliser dans le but auquel il est construit. Dans notre contexte d'étude, ces indicateurs seront utilisés pour réguler la situation d'apprentissage et ajuster le scénario pédagogique, et ceci par la détection des comportements des apprenants observer.

#### 4.2 Modèle d'Indicateur

Le modèle d'indicateur que nous proposons dans la figure 3 s'inspire des modèles étudiés dans certains travaux européens [6][7][9][13][3] (travaux Kaléidoscope, ICALT, IA, TRIAL, DEPLUS, UTL... etc.) .

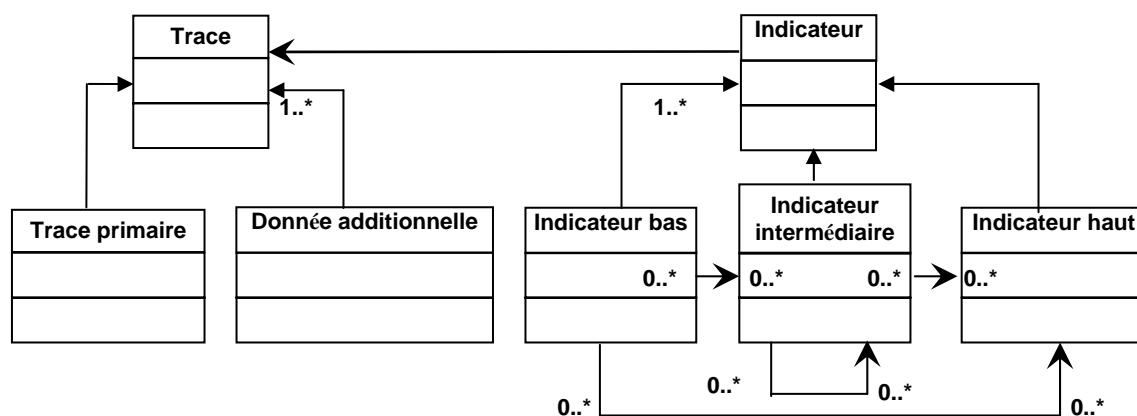


Fig.3. Modèle d'indicateur.

Un indicateur d'apprentissage est une agrégation d'un ou de plusieurs indicateurs et traces. La trace peut être soit une trace primitive, soit une donnée additionnelle (désignée par le concepteur et présente dans la majorité des cas dans le fichier Meta data du cours).

Nous pouvons faire aussi une distinction des indicateurs à travers trois types :

**Indicateur de haut niveau** : il a une interprétation significative. Il est construit par le processus de calcul d'indicateur depuis des données brutes et d'autres indicateurs intermédiaires.

**Indicateur intermédiaire** : sa valeur n'est pas finale mais sert d'intermédiaire pour construire d'autres indicateurs.

**Indicateur de bas niveau** : ce type d'indicateur ne présente pas une interprétation autonome. Il est construit directement par des données brutes.

#### 4.3 Application du Modèle Indicateur au Niveau Scénario

L'apprenant interagit directement avec le système d'apprentissage via une interface. L'unité d'interaction considérée est la page web. Les informations extraites au niveau d'une page (les actions, les durées, l'url... etc.) peuvent être agrégées au niveau activité et au niveau cours. Nous considérons une forme arborescente du scénario d'un cours. Cette forme arborescente a pour racine l'intitulé du cours avec son

objectif pédagogique, composée d'un ou de plusieurs activités, qui peuvent imbriquer d'autres activités. Les activités feuilles pointent vers les ressources que nous considérons des pages.

Afin de comprendre les différents comportements de navigation de l'apprenant et lui adapter le parcours suite à ses interactions, nous examinons une structure de parcours libre.

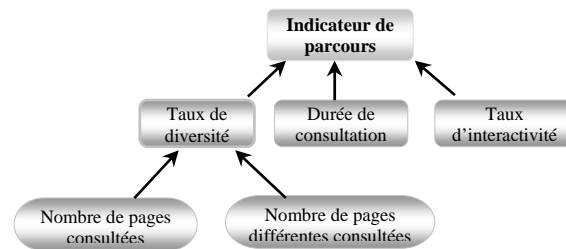
Il est évident qu'il est possible de représenter le scénario comme un graphe de pages dans lequel les comportements des apprenants se traduisent par différents types de parcours.

Pour identifier ces types de parcours, nous nous appuyons ici sur la classification proposée par Canter et al. [1] qui distingue cinq grandes catégories de navigation :

- 1) *Scanning(feuilletage)* : l'utilisateur passe en revue un nombre important de pages sans passer beaucoup de temps sur chaque page, de manière superficielle ;
- 2) *Browsing(navigation)* : l'utilisateur suit un chemin jusqu'à parvenir à son but ;
- 3) *Searching(recherche)* : l'utilisateur cherche un document ou une information en particulier
- 4) *Exploring (exploration)* : l'utilisateur explore une zone ou un domaine particulier jusqu'à en épuiser les ressources ;
- 5) *Wandering (errance)* : l'utilisateur suit un parcours déstructuré et sans but précis.

Obtenir ce type d'information n'est pas aisé car la navigation d'un apprenant est par nature variable. Par exemple, un apprenant badinant sur un cours peut soudainement décider de chercher une information bien précise. De ce fait, il est important d'inférer des informations de haut niveau sur l'apprenant à partir de traces de navigation, relativement de bas niveau (action de recherche, clics, lien vers précédant etc.).

**Calcul d'indicateur.** Pour construire notre indicateur de parcours et identifier les comportements de navigation d'un apprenant, nous avons retenu quelques traces et indicateurs de différents niveaux. La figure suivante (fig4) montre le modèle d'indicateur appliqué à l'indicateur intermédiaire « taux de diversité » pour calculer celui du parcours qui dépend également de la durée de consultation et de taux d'interactivité :



**Fig.4.** Exemple d'indicateur de parcours.

Ainsi pour calculer le taux de diversité (Div) nous appliquons la formule suivante :

$$Div = \frac{Nombredepagesconsultes}{Nombredepagesdifférentesconsultes} \quad (1)$$

La valeur du taux de diversité Div est toujours égale ou bien supérieure à 1. Si la valeur est à 1 nous pouvons dire que le taux de diversité est « neutre » ; il est « moyen » si la valeur ne dépasse pas 2 (en règle chaque page est visitée 2 fois), et « fort » sinon.

La durée de consultation est calculée directement à partir des traces brutes (durée de consultation de chaque page et le nombre de pages consultées).

Enfin, pour calculer le taux d'interactivité d'une action de type x nous calculons d'abord le taux d'interactivité d'une page, ensuite celui d'une activité afin de pouvoir calculer le taux d'interactivité au niveau cours.

*a. Taux d'interactivité d'une page :* Le taux d'interactivité d'une page *p* pour l'action de type *x* est égale au nombre d'actions de type *x* effectuées sur la page *p*, divisé par le nombre totale d'actions effectuées sur toutes les pages du même chapitre père.

$$T_{p_x} = \frac{\sum_{i=1}^N a_{x_i p}}{\sum_{l=1}^M \sum_{j=1}^S a_{j p_l}} \quad (2)$$

$a_{x_i}$  : la  $i^{\text{ème}}$  action de type  $x$  de la page  $p$

$a_{j p_l}$  : la  $j^{\text{ème}}$  action de la page  $p_l$

$M$  : c'est le nombre de pages voisines de la page  $p$  dont nous voulons calculer le taux d'interactivité en plus de la page  $p$ .

$S$  : le nombre total d'actions effectuées sur la page  $p_l$ .

$N$  : le nombre d'actions de type  $x$ , effectuées sur la page  $p$

Si nous désirons calculer le taux d'interactivités de toutes les actions effectuées sur la page  $p$  il suffit de remplacer  $a_{x_i}$  (la  $i^{\text{ème}}$  action de type  $x$ ) par  $a_i$  (la  $i^{\text{ème}}$  action effectuées sur la page  $p$ ) et remplacer le  $N$  par le nombre totale d'actions toujours effectuées sur la page  $p$ . En conséquence nous aurons la formule suivante :

$$T_p = \frac{\sum_{i=1}^N a_i p}{\sum_{l=1}^M \sum_{j=1}^S a_{j p_l}} \quad (3)$$

*b. taux d'interactivité d'une activité*: Le taux d'interactivité de l'action de type  $x$  pour

$$T_{Act_x} = \frac{\sum_{i=1}^N T_{nd_{x_i}}}{N} \quad (4)$$

l'activité  $Act$ , est égale à la somme des taux d'interactivité de l'action de type  $x$  effectuée sur les nœuds fils de  $Act$  divisé par le nombre total de fils de cette même activité  $Act$ .

$T_{Act_x}$  : Taux d'interactivité de l'action de type  $x$  effectuée sur l'activité  $Act$ .

$N$  : le nombre des nœuds fils de l'activité  $Act$ .

$T_{nd_{x_i}}$ : Taux d'interactivité de l'action de type  $x$  effectuée sur le  $i^{\text{ème}}$  fils de l'activité  $Act$ .

*c. Taux d'interactivité d'un cours* : Même formule pour calculer le taux d'interactivité au niveau cours, il est égale à la somme des taux d'interactivité de l'action de type  $x$  effectuée sur les nœuds fils du nœud cours divisé par le nombre total des fils du cours

$$T_{Cours_x} = \frac{\sum_{i=1}^N T_{Act_{x_i}}}{N} \quad (5)$$

$T_{Cours_x}$  : Taux d'interactivité de l'action de type  $x$  effectuée sur le cours.

$N$  : le nombre d'activités fils du cours

$T_{Act_{x_i}}$ : Taux d'interactivité de l'action de type  $x$  effectuée sur la  $i^{\text{ème}}$  activité du nœud cours.

**Déduction des comportements.** De la sorte, nous pouvons justifier les comportements de l'apprenant d'après sa stratégie de navigation en se referant aux indicateurs calculés dans les différents niveaux du cours. Nous pouvons distinguer :

1) *Le comportement de Recherche (Informér)* : ce comportement est retenu lorsque

- Le Taux d'interactivité de type recherche est élevée,
- Le Taux de diversité est fort ( $Div > 1$ ).

2) *Le comportement de découverte (exploration)* : Dans ce cas nous trouvons que :

- La diversité est neutre ( $Div = 1$ ) et
- La durée de consultation est considérable.

3) *Le comportement d'Approfondissement (navigation)* :

- La durée de consultation est considérable (elle peut même dépasser le seuil maximal prévu par le concepteur du cours),

- Diversité moyenne,
- Taux d'interactivité (toujours pour toutes les actions effectuées) élevé.

4) *Le comportement de Survole (feuilletage) :*

- Durée de consultation inférieure au seuil minimal de consultation prévu,
- La diversité est neutre ( $Div = 1$ )
- Le Taux d'interactivité est minimale (tends vers 0)

Le procédé d'adaptation du scénario d'apprentissage, dans notre cas d'étude, est fondé sur une décision pédagogique ou didactique, effectuée de manière manuelle, à partir d'un ensemble d'informations sur l'apprenant. Donc l'objectif c'est d'aider l'enseignant à améliorer son cours selon le comportement des apprenants vis-à-vis du cours consultée (scénario exécutée).

## 5 Conclusion

Le domaine d'analyse des interactions des activités d'apprentissage à distance est une direction de recherche nouvelle, notamment celle des scénarios d'apprentissages. L'objectif de ces travaux est de rendre le système d'apprentissage plus adaptatif et personnalisable pendant son utilisation, et d'offrir une interface de lecture enrichie et une assistance cognitive aux participants et observateurs des systèmes d'apprentissage, par la visualisation d'informations pertinentes. Notre travail s'inscrit dans cette démarche de constitution d'indicateurs pour l'aide à l'adaptation et la personnalisation des apprentissages. C'est ce que nous avons essayé de présenter à travers cet article, en tentant de construire l'indicateur de parcours de l'apprenant par l'analyse de ses interactions avec le système d'apprentissage. Toutefois, notant que l'indicateur ou les indicateurs calculés peuvent être visualisés à l'enseignant pour la régulation et la personnalisation du contenu, à l'apprenant pour une auto-évaluation, ou encore servir d'entrée vers un système d'adaptation automatique.

Nos perspectives proposent de prendre en compte tant le contenu (ressources) que les interactions du cours, d'identifier des méthodes d'analyse des interactions plus puissantes et concevoir des interfaces enrichies, adaptables pour répondre aux exigences distinctes des utilisateurs (enseignant concepteur et tuteur). Il s'agit de créer des interfaces adaptables qui permettent à la fois : d'ajuster ces indicateurs en définissant leurs modèles et une présentation significative de leurs valeurs (visualisation et interprétation).

## Références

1. Canter D., Rivers R., Storrs G. Characterizing user navigation through complex data structures, Behaviour and Information Technology, 4 (2) : 95-102, 1985.
2. Choquet C., Iksal S. Modeling Tracks for the Model Driven Reengineering of a TEL System, Mai 2006.
3. Ferraris C., ans L Vignollet A. L., David J., Modélisation de scénarios d'apprentissage collaboratif pour la classe , EIAH'05, Montpellier, 26 et 27 Mai, 2005.
4. IA: Interaction Analysis. at <http://www.noie-kaleidoscope.org> (2005).
5. ICALTS: Interaction and collaboration analysis Kaleidoscope JEIRP ICALTS, 2004.
6. Lejeune A .,David J-P., Luengo V., Pernin J-P., Diagne F., Adam J-M, Choquet C. State of art of tracking and analysing usage. Délivrable de la tâche 32.3 du projet DPULS .2005.
7. TRAILS Deliverable 4.1 - Personalised Trails and Learner Profiling within e-Learning Environments <http://www.noie-kaleidoscope.org> (31/12/2004).



## Concevoir un tuteur intelligent collectif et adaptatif

HAFIDI Mohamed <sup>1</sup>, BENSBAATaher <sup>2</sup>

Laboratoire de Recherche en Informatique (LRI),  
Département d'informatique – Université Badji Mokhtar,  
BP- 12-23200, Annaba.

E-Mail : <sup>1</sup> [hafidi\\_dz@yahoo.fr](mailto:hafidi_dz@yahoo.fr)  
<sup>2</sup> [t\\_bensbaa@yahoo.com](mailto:t_bensbaa@yahoo.com)

**Résumé.** Les hypermédias permettent de développer rapidement des systèmes interactifs, en particulier dans le domaine de l'éducation. L'hypermédia agit tantôt comme un support de formation (consultation, ...) tantôt comme un support d'enseignement (cours, exercices, ...). Les hypermédias constituent, également, de bons outils pour la représentation des éléments des connaissances des différentes matières. Mais nous savons que les hypermédias présentent deux problèmes, étroitement liés : la désorientation et la surcharge cognitive. Pour palier à ces deux problèmes, nous proposons d'intégrer, à l'hypermédia, un tuteur intelligent capable d'aiguiller l'apprenant désorienté. SAN (Système d'Aide à la Navigation) est un système basé sur une approche statistique et qui cherche à présenter un cours adapté à chaque apprenant en fonction de ses particularités et de ce que l'expérience collective a montré.

**Mots clés :** Surcharge cognitive, Désorientation, Modèle apprenant, Hypermédia adaptatif, Tuteur intelligent, Compréhension, Apprentissage automatique, optimisation.

### 1 Problématique

Dans un vaste espace de connaissances, l'apprenant doit être un acteur actif durant son apprentissage. Il est important pour lui, qu'il puisse distinguer ses forces et ses faiblesses durant tout son parcours pédagogique [5]. Cette forme d'enseignement conduit forcément à une certaine perte dans l'hyperespace [4], qui peut perturber l'apprenant. En effet, plusieurs difficultés peuvent apparaître lors de l'apprentissage des apprenants : surcharge cognitive, désorientation, isolement, etc. Ainsi, il est souvent difficile pour l'apprenant d'identifier des ressources de connaissances pertinentes et de les organiser dans des ensembles cohérents. Nous allons dans cet article, procéder à l'exploration de la problématique relative aux problèmes issus de l'utilisation des hypermédias [3] à des fins pédagogiques.

### 2 Elaboration de la solution

Notre solution repose sur l'utilisation d'une approche statistique pour offrir une personnalisation plus importante de l'enseignement à l'apprenant. Notre démarche s'appuie sur le fait de caractériser et de profiler les apprenants pour les faire profiter des expériences déjà acquises par d'autres apprenants au même profil.

Ainsi nos questions portent sur les points suivants :

- Comment détecter l'état de l'apprenant ?
- Comment arriver à guider l'apprenant en cas de désorientation ?
- Comment réutiliser les expériences des autres apprenants ?

#### 2.1 Choix de l'approche statistique

Traditionnellement, la conception des systèmes tuteurs est basée sur un ensemble de règles pédagogiques élaborées par l'auteur ou le concepteur du système. Ces règles sont définies et développées manuellement en se basant sur l'expérience de l'auteur du système tuteur. Cette logique « manuelle » est plus subjective qu'objective car elle est liée à l'auteur et à ses expériences pédagogiques. Notre idée est de remplacer cette approche par une approche statistique. Il faut développer des approches probabilistes et des modèles flexibles, qui permettent à la machine d'apprendre par elle-même à partir des données reçues. Notre approche rassemble de l'information au sujet du comportement de l'apprenant à chaque parcours de l'apprentissage individualisé afin de créer une distribution des probabilités pour les autres apprenants.

## 2.2 Indice de désorientation

Pour détecter un apprenant désorienté, nous nous basons sur quatre variables :

- *La redondance (red)*: l'apprenant peut consulter l'historique pour rappeler un concept. Mais le problème se pose si ce dernier demande la même page plusieurs fois.
- *Le nombre de lectures supplémentaires (lect-supp)* : l'apprenant peut demander une lecture supplémentaire pour comprendre le concept en cours. Il y a problème si cette demande est répétée plusieurs fois.
- *Le nombre d'exercices supplémentaires (exo-supp)* : dans un cas d'échec, l'apprenant peut demander un exercice supplémentaire.
- *Le nombre de retours à une partie antérieure (ret)* : le problème se pose si l'apprenant a acquis un chapitre et demande, ultérieurement, le retour à ce chapitre.

Nous considérons comme *indice d'une désorientation* la somme des différentes variables :

*Indice d'une désorientation (D)=red+lect-supp+exo-supp+ret.*

## 2.3 Guidage de l'apprenant

Pour chercher l'enchaînement de concepts sur lequel l'apprenant maximise son apprentissage ainsi que sa probabilité de succès, nous proposons de valuer les arcs du graphe représentant l'hypermédia. Il est ainsi possible d'estimer, à un instant donné, le taux de réussite des apprenants quant à leur passage du nœud source au nœud destination. Comme le succès (échec) sur un nœud donné est conditionné notamment par ce que l'apprenant a vu et fait avant, alors la solution cherchée n'est pas seulement un lien entre deux nœuds, mais plutôt un chemin pédagogique et donc un succès (échec) de liens visités.

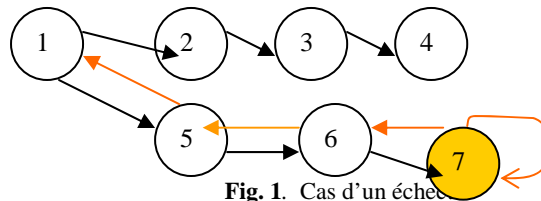


Fig. 1. Cas d'un échec

### 2.3.1 Construction de la fonction d'orientation

Notre objectif est de déterminer le chemin à emprunter par l'apprenant, lorsque celui-ci veut changer de nœud. Pour cela, nous devons déterminer le lien à emprunter.

Pour ce faire, nous attribuons à chaque lien une valeur, appelée fonction d'orientation, qui est définie comme suit :

$$F=(PCW+PCS-PCE)*PIH*PIR*PIM*PIC.$$

Où :

- PIH (paramètre individuel historique) : porte l'information des nœuds précédemment visités par l'apprenant. Son rôle est de diminuer la probabilité que le nœud déjà visité soit à nouveau proposé.
- PIR (paramètre individuel révision) : pousse l'apprenant à aller vers certains nœuds nommés nœuds critiques. Soit le système juge nécessaire d'y passer, soit qu'ils aient été conseillés par un enseignant.
- PIM (paramètre individuel module) : focalise l'apprenant sur un même module.
- PIC (paramètre individuel chapitre) : focalise l'apprenant sur un même chapitre.
- PCS (paramètre collectif succès) : utilisé pour représenter le nombre de succès.
- PCE (paramètre collectif échec) : utilisé pour représenter le nombre d'échecs.
- PCW (paramètre collectif poids pédagogique) : établi par des enseignants en fonction de leur vision du module. Plus le poids est important, plus l'arc concerné est avantagé pour être sélectionné.

### 2.3.2 Paramètres de la fonction d'orientation

#### a) Paramètre individuel historique (PIH)

Il existe une valeur H (initialement égale à 1) par apprenant et par nœud. Lorsqu'un nœud est visité par un apprenant, la valeur H correspondante est multipliée par :

- 0.25 s'il s'agit d'un succès
- 0.75 s'il s'agit d'un échec.

Si le nœud a été le siège d'un échec, il sera plus vite proposé à nouveau.

Ainsi si l'apprenant réussit un concept, et que par la suite il a de nouveau la possibilité de revenir sur le concept en question, la valeur des arcs y menant est alors divisée par deux. Ceci signifie que ces arcs ont deux fois moins de chances d'être sélectionnés par le système pour être proposés à l'apprenant.

#### **b) Paramètre individuel révision (PIR)**

A chaque fois qu'un concept est désigné par le système en tant que concept intéressant pour un apprenant particulier, le PIR associé à chaque apprenant et au concept, est incrémenté d'une certaine valeur. Lorsque la valeur du PIR dépasse un certain seuil, le concept est jugé critique. Ce facteur ne peut revenir à 1 qu'une fois le concept concerné est acquis.

#### **c) Paramètres individuels module et chapitre (PIM, PIC)**

Pour focaliser l'apprenant sur un même chapitre (même module), il faut mettre des bornes afin de ne pas proposer des liens trop excentriques. La priorité est donnée au module et chapitre en cours.

$$PIM = \begin{cases} 1 & \text{si } Ma <> Mb \\ 2 & \text{si } Ma = Mb \end{cases}$$

Si l'apprenant sélectionne un concept qui appartient au module Mb différent du module Ma alors le paramètre PIM prend la valeur 1.

$$PIC = \begin{cases} 1 & \text{si } Ca <> Cb \\ 2 & \text{si } Ca = Cb \end{cases}$$

Si l'apprenant sélectionne un concept qui appartient au chapitre Cb différent du chapitre Ca alors le paramètre PIC prend la valeur 1.

#### **d) Paramètres collectifs succès et échec (PCS, PCE)**

Si l'apprenant réussit un exercice, le système utilise le paramètre de succès PCS pour représenter ce succès.

Si par contre, l'apprenant a échoué, le système utilise le paramètre PCE pour représenter cet échec.

#### **e) Paramètres collectifs poids pédagogique (PCW)**

Un module est modélisé suivant un graphe où chaque nœud représente un élément pédagogique (leçon, tp, exercice). Chaque arc, représentant une navigation possible entre deux éléments, porte le poids pédagogique PCW. En revanche, l'apprenant n'est pas obligé de suivre la proposition faite par le tuteur. S'il le souhaite, il peut se diriger vers un nœud non proposé. Si l'arc alors emprunté n'existe pas, il est créé et son poids pédagogique PCW est fixé à 1. Un tel arc est appelé « arc apprenant » par opposition aux arcs enseignants ( $PCW > 1$ ).

### **2.3.3 Construction de la fonction de sélection**

Une fois le chemin établi, le système commence à indexer et filtrer les documents pour une meilleure présentation.

Pour cela, il se base sur une fonction de sélection d'un document :

$$F_{select} = F_{pres} * F_{faut} * PIN(S - E + D)$$

Où :

- S: nombre de succès.
- E : nombre d'échecs.
- D : degré de difficulté d'un document.
- PIN : niveau de l'apprenant.

$$- \text{Faut} = \begin{cases} 1 & \text{si AUTa} <> \text{AUTb} \\ 2 & \text{si AUTa} = \text{AUTb} \end{cases}$$

*Tel que* : AUTa est l'auteur préféré par l'apprenant.  
AUTb est l'auteur du document.

$$- \text{Fpres} = \begin{cases} 1 & \text{si PRESa} <> \text{PRESb} \\ 2 & \text{si PRESa} = \text{PRESb} \end{cases}$$

*Tel que* : PRESa est la présentation préférée par l'apprenant.  
PRESb est la présentation du document.

### 3 Architecture du système SAN

SAN est constitué des composants standards d'un hypermédia adaptatif dynamique [1] [2] en plus d'un tuteur intelligent [7] [8] : un module de représentation de la connaissance du domaine, un modèle de l'apprenant, une base multimédias pour les documents et un tuteur pour évaluer, conseiller et orienter l'apprenant désorienté.

Le principe de fonctionnement de notre système est le suivant :

- Le tuteur commence par l'initialisation des paramètres du système.
- L'apprenant choisit un module d'enseignement, le tuteur expose le contenu de ce module, chapitre par chapitre. Le passage d'un chapitre à l'autre est conditionné par le contrainte que le taux de réussite est supérieur à un seuil, quelque soit le niveau de l'apprenant.
- Suite à une demande de l'apprenant « navigation guidée », ou en cas de blocage détecté par le tuteur, ce dernier propose un nœud. Ceci est fait par le calcul d'une fonction d'orientation pour chaque arc sortant du nœud courant.
- Dans le cas où le nœud sélectionné est de type cours ou TP, le tuteur applique une fonction de sélection sur les documents associés à ce nœud, pour choisir le meilleur d'entre eux.
- Dans le cas où le nœud sélectionné est de type exercice, le tuteur récupère la page d'exercice correspondant au niveau de difficulté choisi par l'apprenant.
- Dans le cas où l'apprenant réussit, il a donc la possibilité soit de changer de niveau d'exercice, ou bien de poursuivre l'apprentissage.
- Dans l'autre cas (échec), l'apprenant a la possibilité soit de demander un exercice supplémentaire, soit de changer de niveau d'exercice, ou bien de consulter les concepts critiques.
- Pour chaque scénario d'évaluation, le tuteur met à jour les paramètres collectifs des arcs visités et les documents consultés, ainsi que les paramètres individuels du modèle apprenant.

#### 3.3 Modèle du domaine

Le modèle du domaine est un ensemble de modules [9] [10] [11] [125].

Un module est une séquence de chapitres. Un chapitre est structuré en Meta concepts et se termine par une série d'exercices. Un meta concept est composé d'un concept de type cours et un ensemble de concepts de type TP et se termine par un ou plusieurs concepts de type exercice (Fig 2). Un chapitre est modélisé suivant un graphe où chaque nœud représente un élément pédagogique (cours, TP, exercice). Chaque arc, représentant une navigation possible entre deux éléments, porte les paramètres collectifs.

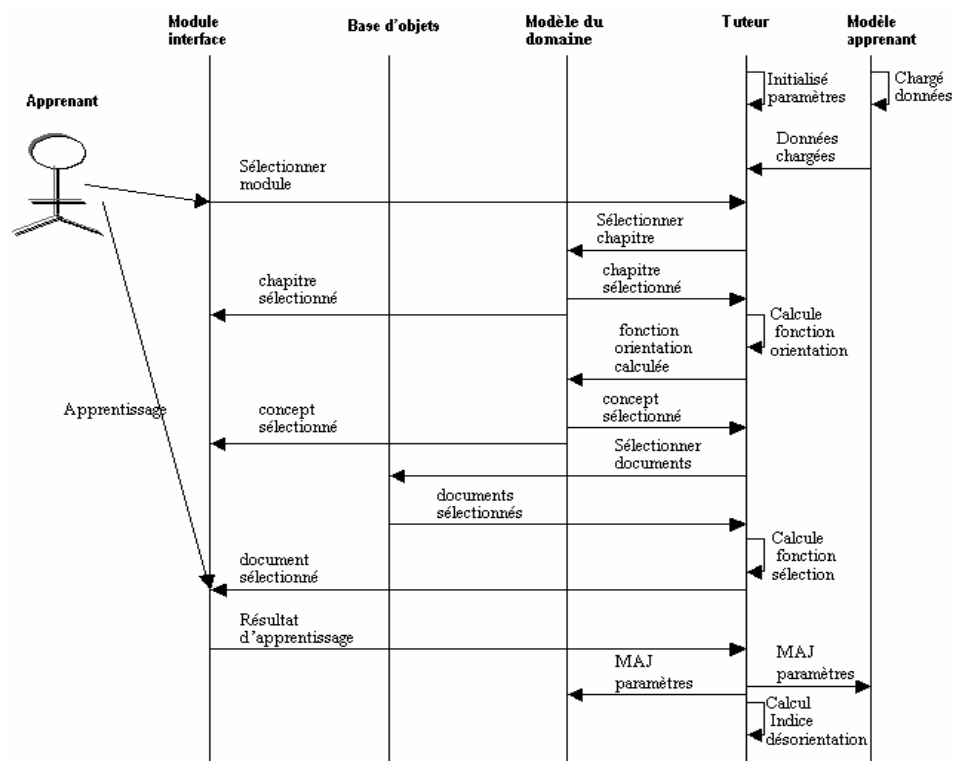
Un concept est une entité pédagogique, qui comporte plusieurs types :

##### 3.1.1 Les concepts de type cours :

Constituent des rappels de cours. Il en existe plusieurs variantes selon le profil de l'apprenant.

##### 3.1.2 Les concepts de type Tp :

Représentent les méthodes utiles à connaître pour une bonne utilisation des notions de cours. C'est la version pratique du cours. Ces concepts constituent une passerelle entre les concepts de cours et les exercices.



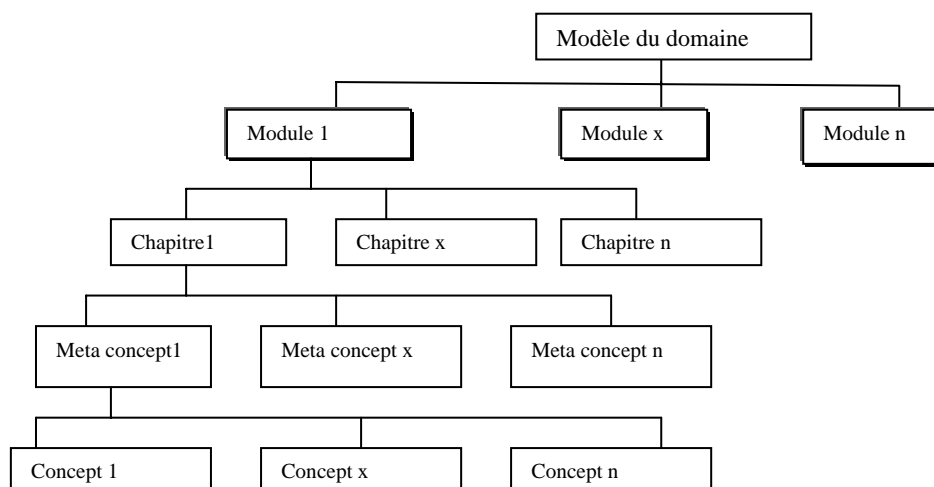
**Fig 2.** Les scénarios d'interactions

### 3.1.3 Les concepts de type exercice :

Un exercice est caractérisé par un niveau de difficulté affecté au départ par l'enseignant.

Le niveau de difficulté d'un exercice peut évoluer au cours du temps selon son taux de réussite. Si beaucoup d'apprenants échouent sur un exercice, cela signifie que l'exercice proposé est assez difficile et son niveau peut par conséquent être revu à la hausse. Inversement si les apprenants réussissent bien un exercice, son niveau peut être revu à la baisse. Voici comment nous avons géré ce fait :

$$\text{Taux de réussite d'un exercice} = \frac{(\text{Nombre de fois qu'un exercice est réussi})}{(\text{Nombre de fois qu'un exercice est proposé})}$$



**Fig 3.** L'architecture du domaine d'enseignement

### 3.2 Modèle apprenant

Pour proposer des cheminements de cours différents, nous avons cherché des critères permettant de différencier les apprenants dans leur apprentissage.

Nous avons alors défini le profil d'un apprenant : il s'agit de caractéristiques personnelles qui permettent de retrouver dans les cours ce qui convient le mieux à cet apprenant.

Actuellement, le modèle comporte :

- Une préférence de présentation
- Une préférence de contenu.
- Un niveau de difficulté.
- L'auteur préféré.
- Modes d'apprentissage.

### 3.3 Le tuteur

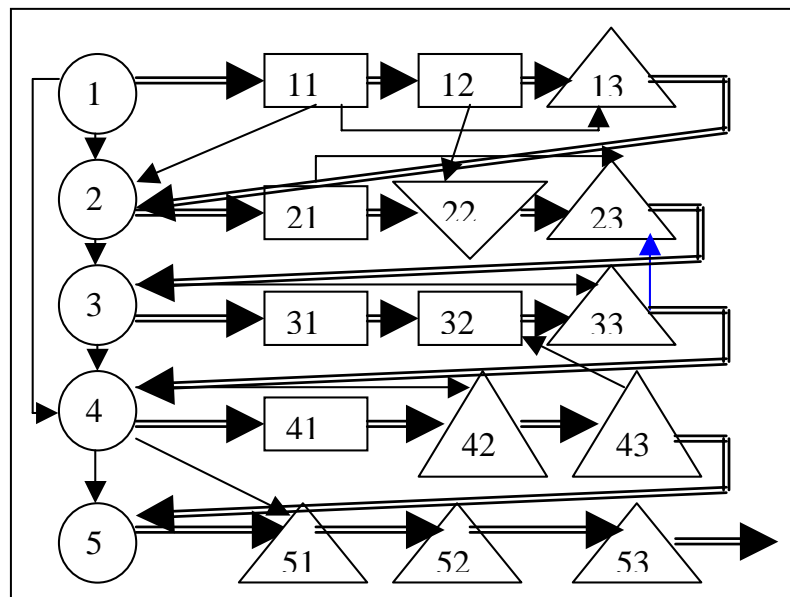
Le rôle du tuteur dans le système SAN est de guider, conseiller l'apprenant désorienté. Il repose sur deux piliers. Le premier est l'adaptation d'une fonction d'orientation qui suggère un chemin. Le second est l'utilisation d'une fonction de sélection qui donne la meilleurs page à proposer à l'apprenant.

Ces fonctions utilisent deux types de paramètres :

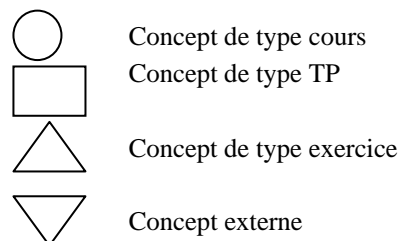
- Paramètres individuels servant à aiguiller l'apprenant sur un des nœuds accessibles en fonction de son profil.
- Paramètres collectifs afin de l'aiguiller en fonction de l'efficacité de ces nœuds.

## 4 Résultat et discussion :

Nous avons pris comme exemple un cours d'analyse de données.



**Fig 4.** Exemple d'un chapitre



Le chapitre ACP (analyse en composantes principales) est structuré en les meta-concepts suivants : présentation de la méthode, représentation des individus, représentation des variables, interprétation et conclusion.

Le meta concept interprétation est composé de :

- 4 : inertie
- 41 : contribution
- 42 : interpretation1
- 43 : interpretation2

Le meta concept conclusion est composé de :

- 5 : conclusion générale
- 51, 52,53 : série d'exercices.

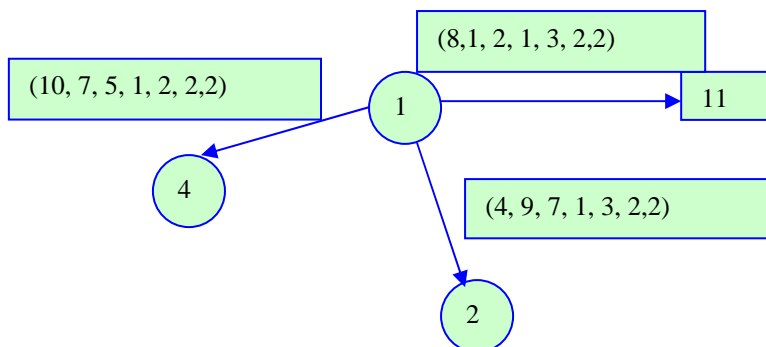
L'analyse de ce chapitre va permettre de mieux comprendre le principe de fonctionnement du tuteur. Il traite l'ACP, il possède 20 concepts dont 5 de type cours, 6 de type TP, 8 de type exercice et un concept externe (appartient à un autre module). Il présente 5 Meta concepts identifiés chacun par un concept de type cours suivi par une série de TP et d'exercices :

- Présentation de la méthode : (1) de type cours, (11,12) de type TP, (13) de type exercice.
- Représentation des individus : (2) de type cours, (21) de type TP, (22) externe, (23) de type exercice.
- Représentation des variables : (3) de type cours, (31,32) de type TP, (33) de type exercice.
- L'interprétation : (4) de type cours, (41) de type TP, (42,43) de type exercice.
- Conclusion : (5) de type cours, (51, 52,53) de type exercice.

La figure représente la navigation d'un apprenant A (Faut, Fpres, PIN, PIA) qui débute sur le concept (1) et suit à chaque fois l'arc possédant la meilleure fonction d'orientation.

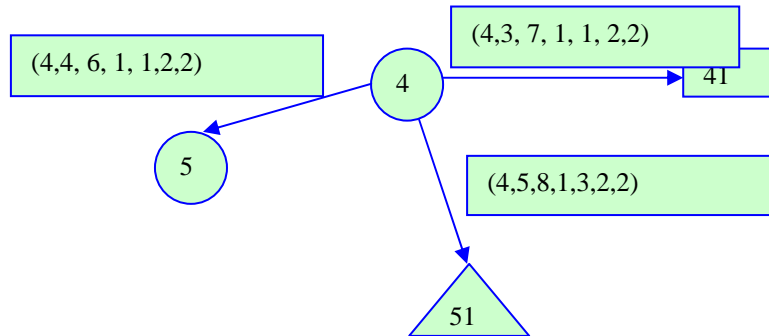
Depuis le concept (1), le tuteur peut suggérer d'aller soit vers le concept (2), soit vers le concept (11) ou bien vers le concept (4).

Par exemple : pour un apprenant A (akram, Txt, débutant, cours), et le graphe G(PCS, PCE, PCW, PIH, PIR, PIC, PIM):



road	pcs	pce	pcw	pih	pir	pic	pim	Orientation
4	10	7	5	1	3	2	2	96
11	8	1	2	1	1	2	2	36
2	4	9	7	1	1	2	2	8

Le tuteur suggère le concept (4) où la fonction d'orientation est de 96.



nœud	pcs	pce	pcw	pih	pir	pic	pim	Orientation
5		4	4	6	1	1	2	24
41		4	3	7	1	1	2	32
51		4	5	8	1	3	2	84

Le tuteur suggère le concept (51) où la fonction d'orientation égale 84.

Ces résultats nous permettent de dire que :

- l'introduction des paramètres dans le système permettent à l'apprenant de sauter quelques concepts. En l'absence de ces paramètres, le tuteur propose le chemin pédagogique présenté dans la figure par le double flèche (  $\Rightarrow$  ).
- Le concept (4) qui a été proposé par le tuteur, est un concept critique et facile.
- Le concept (51) est un concept critique et très important (poids pédagogique important).
- Le concept (2) est un concept difficile car le nombre d'échecs est très important.
- etc.

## 5. Conclusion et Perspectives

Dans cet article, notre souci est de palier l'absence de l'enseignant par l'utilisation d'un système d'accompagnement (SAN) afin d'orienter l'apprenant désorienté. Notre démarche s'appuie sur le fait de caractériser et de profiler les apprenants pour les faire profiter des expériences déjà acquises par d'autres apprenants au même profil. L'originalité du système SAN réside dans le paramétrage et l'approche statistique que nous avons adoptée d'une part, l'introduction des paramètres collectifs et les variables de désorientation d'autre part. Les principales contributions de SAN sont :

- la capacité à détecter les incohérences lorsqu'elles surviennent (adaptabilité,...), ou les erreurs de conception (mauvais poids pédagogique,...).
- Proposer des exercices difficiles aux apprenants, leur permettant de progresser, tout en leur laissant une impression de succès pour les motiver.
- La possibilité de restructurer le chapitre en mettant le Meta concept qui possèdent les concepts les plus simples en premier, afin de mettre les apprenants en confiance.
- Evoluer le profil de l'apprenant s'il réussit mieux sur des pages de cours qui ne correspondent pas à son profil.

Durant cette première version, on a pu dégager quelques problèmes comme le temps de réponse est très important, absence des possibilités de réutiliser le raisonnement des autres apprenants, la possibilité d'ajouter ou de supprimer des paramètres afin d'améliorer l'adaptation, la possibilité de proposer d'autres variables de désorientation, absence une priorité entre les paramètres individuels et les paramètres collectifs. Ces critiques sont importantes et sont prises en compte dans une version améliorée du système SAN, pour une deuxième version avec un cours sur l'analyse de données pour les apprenants du MASTER et avec un échantillon plus grand.



## 6. Références

- [1] N. Delestre, « Un hypermédia adaptatif dynamique pour l'enseignement », Thèse au laboratoire PSI de l'université de Rouen, 2000.
- [2] P. Brusilovsky, « Methods and techniques of adaptive hypermedia », Adaptive hypertexte and hypermedia, Kluwer Academic Publishers, p 1-43, 1998.
- [3] A. Tricot, F. Amadieu, « Utilisation d'un hypermédia et apprentissage : deux activités concurrentes ou complémentaires ? », Revue Psychologie Française, 2006.
- [4] S. Chabert-Ranwez, « Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur », thèse de doctorat de l'Université de Montpellier II, 2000.
- [5] J. Labat, « EIAH: Quel retour d'information pour le tuteur ? », Actes du colloque Technologies de l'information et de la Communication dans les enseignements d'ingénieurs et dans l'industrie Lyon, 2002.
- [6] S. Benadi, « Structuration des données et des services pour le téléenseignement », Thèse de doctorat de l'Institut National des Sciences Appliquées de Lyon, 2004.
- [7] M. Hafidi, T. Bensbaa, « Intégration d'un tuteur intelligent dans un hypermédia adaptatif dynamique, une approche fondée sur des ontologies », COSI, Bejaia, Algérie, Juin 2005.
- [8] M. Hafidi, T. Bensbaa, « Intégration d'un tuteur intelligent dans un hypermédia adaptatif dynamique, une approche fondée sur des ontologies », CEMAFORAD 2, Bejaia, Algérie, Nov 2005.
- [9] M. Hafidi, T. Bensbaa, « Integration of multituteur ontology based on the dynamic adaptive hypermedia (MODAH) »; multi séminaire csit'06, amman, Jordan, 2-4 avril 2006, page 163, vol 2.
- [10] M. Hafidi, T. Bensbaa, « Une architecture multituteur dans un hypermédia adaptatif dynamique », conférence internationale sur l'informatique et ses applications ciia'06, saida 14-16 mai 2006.
- [11] M. Hafidi, T. Bensbaa, « Implantation of multituteur ontology based on the dynamic adaptive hypermedia (MODAH) », CEMAFORAD 3, sousse, Tunisie, nov, 2006.
- [12] M. Hafidi, T. Bensbaa, « Proposition d'un modèle du multitutorat pour le suivi des apprenants », SITA'06, el mohamadia, Maroc, nov, 2006.

## Processus d'adaptation des activités et des documents dans un cours en ligne

Bekkaoui Mokhtaria<sup>1</sup>, Chikh Azzeddine <sup>1</sup>

<sup>1</sup> Université Abou Bekr Belkaid - Tlemcen  
B.P.230- Tlemcen 13000 Algérie  
[{m\\_mekkaoui, a\\_chikh}@univ-tlemcen.dz](mailto:{m_mekkaoui, a_chikh}@univ-tlemcen.dz)

**Abstract.** L'apprentissage est le processus d'acquisition de connaissances, compétences, attitudes ou valeurs, par l'étude, l'expérience ou l'enseignement. La problématique de la conception des situations d'apprentissage adaptatives par l'enseignant occupe une place stratégique dans le champ des EIAH. Nous proposons un double processus d'adaptation relatif respectivement aux activités et aux ressources. Ce processus utilise une comparaison des métadonnées extraites à partir des schémas LOM, IMS-LD et IMS-LIP qui qualifient respectivement les ressources, les activités et les apprenants.

**Keywords:** E-Learning, Adaptation, SHA, Objet pédagogique, Scénario d'apprentissage, Activité, LOM, IMS LD, IMS LIP.

### 1 Introduction

E-Learning. Celui-ci est basé sur l'accès à des formations en ligne, interactives et parfois personnalisées, diffusées par l'intermédiaire d'un réseau (Internet ou Intranet) ou d'autres média électroniques. Cet accès permet de développer les compétences des apprenants, tout en rendant le processus d'apprentissage indépendant du temps et du lieu. Ce développement conduit à utiliser la notion d'adaptation ; c'est un mécanisme qui vise à répondre de façon adaptée aux besoins et caractéristiques uniques et particuliers de chaque apprenant. Dans cette nouvelle direction de recherche, nous nous intéressons au mécanisme d'adaptation d'un scénario d'apprentissage aux profils de l'apprenant.

Dans une première section, nous allons mettre le point sur la notion d'adaptation dans le contexte de E-Learning en étudiant les différents projets qui traitent cette notion. La deuxième section présente notre processus proposé pour l'adaptation d'un scénario d'apprentissage par rapport aux caractéristiques et préférences des différents apprenants. Enfin, pour conclure, nous traçons les perspectives de recherche ouverte par nos propositions.

### 2 L'adaptation en e-Learning

L'adaptation dans le contexte de e-Learning vise à créer l'expérience de l'apprenant sur une certaine période de temps. Il y a beaucoup de projets qui traitent la notion d'adaptation sur les deux axes, le scénario d'apprentissage et l'objet pédagogique. Ils sont différents les uns par rapport aux autres selon différentes facettes, par exemple : méthode et technique d'adaptation, modèle d'apprentissage, situation d'apprentissage, les caractéristiques d'apprenant et standard de e-Learning.

Dans les projets qui traitent la notion d'adaptation sur le scénario d'apprentissage, notre analyse est faite en prenant en considération le type d'adaptation, dimension d'adaptation et critères d'adaptation. Nous avons remarqué que les solutions proposées basées soit sur l'adaptation statique c'est le cas de projet Alfanet [1], soit sur l'adaptation dynamique tel que le projet Explora Graph [7] et le travail de l'équipe arcade [2]. Pour la dimension d'adaptation, il y a ceux qui s'intéressent à l'adaptation de contenu tel que le travail de l'équipe arcade et le projet Alfanet et ceux qui sont basés sur l'adaptation de navigation c'est le cas de projet Explora Graph. Pour les critères d'adaptation, les différents projets ci-dessus ne traitent pas toutes les caractéristiques d'un apprenant à la fois, ils sont basés sur la connaissance de l'apprenant.

Dans les projets qui traitent la notion d'adaptation sur l'objet pédagogique, notre analyse a été faite en prenant en considération le modèle de document où il y a les deux notions pour représenter un document « Brique Élémentaire » ou « Page ».

Plusieurs travaux ont été réalisés dans le domaine de la modélisation des documents électroniques à des fins diverses.

Dans METADYNE [5] une granularité plus fine que les Brique d'Information a été adoptée, il s'agit des Briques Élémentaires. Une brique élémentaire se définit comme l'élément de base qui compose le document. Il peut être un élément textuel, une image, un schéma, une animation, etc. le document est un assemblage des BE qui sont sélectionnées soit par l'auteur soit, dans d'autres situations, par le système.

Dans EDPHA [10], le document est vu comme un ensemble de pages (au sens des pages web). Chaque page est découpée en blocs de données. Un bloc est une collection d'éléments dont chacun est une description du même concept (définition, exemple, description, exercice, etc.) à partir de divers médias (« texte+image+son », « texte+son », « animation+son »). Les éléments, sont représentés par une collection de médias tels qu'une image avec un texte (ce qui correspond par exemple à une image avec sa légende). Il s'agit de regrouper les données indissociables, c'est pour cela que la granularité n'est pas la plus fine, en l'occurrence les médias.

L'analyse présentée ci-dessus, aboutit au constat que les projets étudiés sont fortement centrés sur l'adaptation de contenu ou navigation par rapport à la connaissance de l'apprenant.

Partant de ces constats, nous situons donc notre travail dans la mise à disposition d'un environnement adapté aux apprenants, permettant la modification dynamique d'un scénario d'apprentissage, et prenons en considération l'avantage de l'approche centré sur le contenu et l'approche centré sur le processus.

#### 4 La modélisation du processus d'adaptation

Les formations auxquelles nous nous intéressons principalement sont des formations de type e-Learning. Plus précisément, adoptant un point de vue constructiviste, nous considérons que l'apprenant doit avoir un rôle actif dans sa formation. Le contenu mis à sa disposition n'est pas un simple cours en présentiel. Il s'agit au contraire d'un *scénario pédagogique* : le scénario prévoit le déroulement d'une activité d'apprentissage et comprend une définition des objectifs, une planification des activités, une description des tâches des apprenants et des modalités d'évaluation.

Nous voulons par notre travail mieux soutenir l'apprenant. nous nous intéressons à la dimension d'adaptation utilisée dans la taxonomie de Dieterich [6] [10]. Nous étudions l'adaptation dynamique 'adaptativité initiée par l'apprenant' qui va être modélisé par la construction d'une interface interactive avec l'apprenant où elle permet la capture et le traitement des requêtes qui agissent sur le déroulement du scénario pour mettre à jour les profils des apprenants tel que : préférences, objectifs.

Dans notre analyse nous abordons une question importante ' Est-ce que la 'method' est réutilisable dans un autre contexte ? '

Considérons la remarque suivante : « Grâce à l'indépendance des composants, IMS LD favorise grandement la réutilisation des composants dans d'autres 'Learning Design'. Le seul composant difficilement réutilisable est la section 'method' parce qu'il crée le lien entre tous les autres [14] », donc au lieu d'utiliser 'method' nous proposons d'utiliser les morceaux suivants : Activité et Ressource.

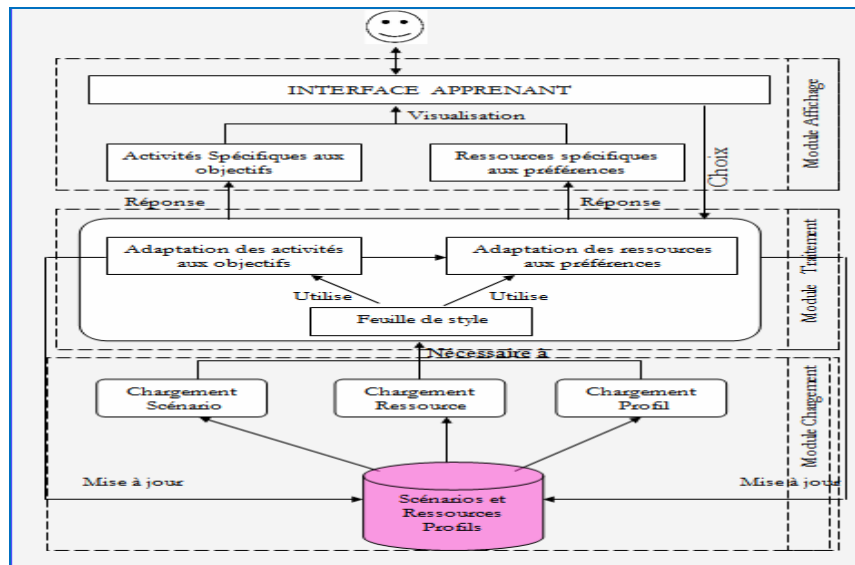
A la lumière de cette remarque, notre démarche est partagée en deux parties successives :

La 1<sup>ère</sup> étape : Etablir un mécanisme d'adaptation de contenu des activités dans une situation d'apprentissage, via l'interaction du système avec l'apprenant, qui nous amène à une scénarisation dynamique nous permettre la mise à jour d'un scénario prédictif, dans cette partie l'adaptation basée sur IMS LD prend en charge les objectifs d'un apprenant. Cette adaptation permet aux apprenants de suivre un parcours adapté à leur formation antérieure.

Nous utilisons la notion 'Adaptabilité' au niveau de scénario prédictif, tel que le scénario est conçu satisfaisant les besoins d'un apprenant générique. Le choix de scénario fait selon ce que l'auteur du cours a spécifié comme caractéristique d'un apprenant au niveau macro.

La 2<sup>ème</sup> étape : Etablir un mécanisme d'adaptation de navigation des ressources manipulés par les activités, nous prenons en charge les préférences de l'apprenant. Cette adaptation a pour objectif de donner à l'apprenant une certaine flexibilité dans l'utilisation de ces documents. Cette méthode est vue comme une solution pour deux risques majeurs pour l'utilisateur : la désorientation et le phénomène de surcharge cognitive. A ce niveau, l'adaptation est basé sur le modèle LOM.

La figure.1 présente l'architecture globale du processus d'adaptation.



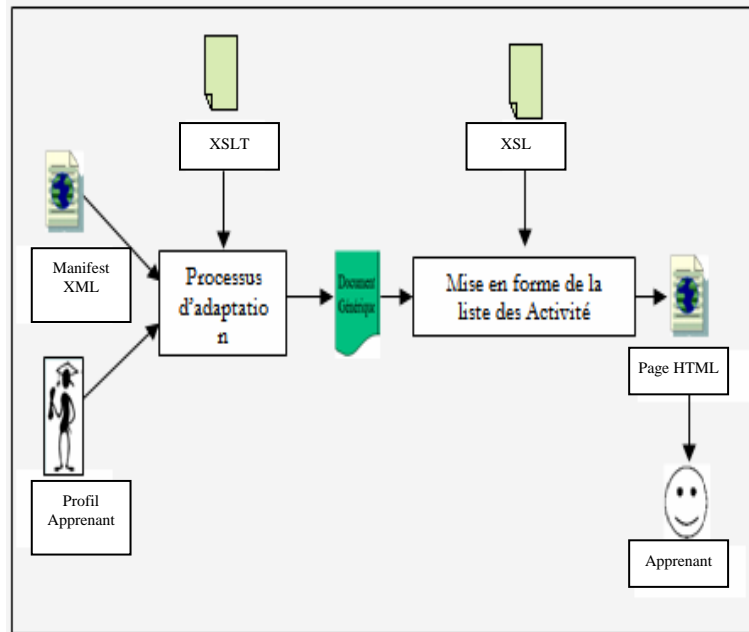
**Fig .1.** Architecture générale du processus d'adaptation

Notre système est décomposé en trois modules sur lesquels se base le processus d'adaptation du scénario.

1. *Le module de chargement* : ce module regroupe une base de données, parmi lesquelles nous distinguons celle qui est lié aux scénarios et ressources, ainsi que les informations liées aux apprenants. Via ce module, le système permet de charger, sauvegarder les différents scénarios, les ressources et les caractéristiques des apprenants au format XML.
2. *Le module de traitement* : c'est le processus intelligent qui produit l'adaptation. Il consiste à confronter les données et exécuter en appliquant les règles d'adaptation, il est basé sur la sélection des activités qui soient les plus compatibles avec les objectifs de l'apprenant dans la première étape. Pour la deuxième étape ça sera un ensemble de choix effectué par l'apprenant afin de lancer une recherche pour sélectionner un ensemble de briques.
3. *Le module affichage* : Ce module regroupe un ensemble organisé d'objets interactifs définissant la partie perceptible du système (boutons, curseur, barre de défilement, menus déroulants, etc.). L'objectif de module, est de rendre le système facile à manipuler en présentant à l'apprenant une vue explicite qui ne comporte pas d'ambiguïtés. A part les menus déroulant traduisant les différentes fonctions, nous trouvons une boîte de dialogue qui traduise les préférences de l'apprenant.

#### 4.1 Le processus d'adaptation des activités

Au niveau de modélisation, le modèle IMS LD est composé de trois niveaux A, B et C, dans notre étude nous allons nous baser sur le niveau B où nous utilisons les propriétés pour spécifier les conditions d’affichage d’une activité d’après les objectifs de l’apprenant. Le processus d’adaptation peut être schématisé comme suit :



**Fig .2.** Adaptation aux objectifs d'un apprenant

Nous avons trois entrées utilisées dans le processus ; le *manifest XML* qui décrit le scénario, Le *profil apprenant* ; un fichier XML décrit les profils d'apprenant

Le choix des activités qui vont être affichées à l'apprenant est réalisé par des règles XSLT c'est la troisième entrée qui va être appliquée au moment du déroulement de l'unité d'apprentissage. Pour satisfaire les objectifs de l'apprenant, nous écrivons un programme script qui permet de faire le 'Matching' entre le scénario d'apprentissage 'manifest.xml' et le profil de l'utilisateur modélisé par IMS LIP 'Profil.xml'.

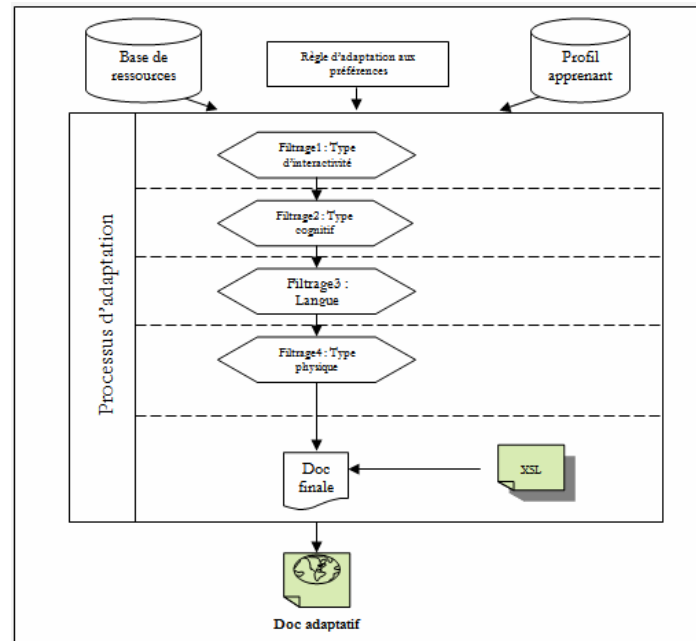
#### 4.2 Le processus d'adaptation des ressources

Une fois l'activité est choisie, le système donne la main à l'apprenant de sélectionner un ensemble des critères de recherche. Ces derniers fournissent le choix effectué par l'apprenant. Cette sélection nous donne les critères de filtrage dans le modèle du domaine. Pour ce dernier, nous adoptons la granularité proposé dans le projet METADYNE [5] ; une Brique Élémentaire (BE) se définit comme l'élément de base qui compose le document. C'est à partir de ce travail que nous représentons le modèle de ressource comme suit :

Le concept de BE est caractérisé par quatre attributs :

- Type physique : représente l'aspect physique de la brique.
- Type d'interactivité : montre si la brique nécessite une interaction avec l'apprenant ou non.
- Type cognitif : représente l'aspect cognitif de la brique.
- Langue : représente les niveaux de maîtrise une langue pour utiliser la brique.

Pour avoir la ressource « BE » qui répondra mieux aux préférences de l'apprenant, il y a une certaine étape de filtrage comme le montre la figure 5



**Fig .5.** Le mécanisme d'adaptation aux préférences d'apprenant

Dans notre cas, le processus d'adaptation est basé sur la sélection d'une combinaison des briques qui soit la plus compatible avec le profil d'apprenant.

Comme tout système, nous avons comme entrée les ressources incluent dans un scénario pédagogique et le profil d'apprenant. L'adaptation opérée sur les ressources nécessite une application des règles XSLT générées automatiquement.

Pour traiter les préférences, nous procédons par une démarche de filtrage progressif des BE alternatifs qui va être réduit au fur et à mesure jusqu'à la sélection de l'objet pédagogique le plus adapté à l'apprenant.

Etape1 : la première étape de filtrage consiste à extraire un ensemble de briques élémentaires en fonction de type d'interactivité 'Actif, Expositif'

Etape2 : il s'agit d'effectuer la même opération mais pour un type cognitif déterminé 'définition, Exercice, etc.'

Etape3 : l'ensemble des composants qui restent, seront soumis à une adaptation au langage. Nous tenons à rappeler, ici, que le modèle LOM manque de flexibilité qui prenne en considération la langue de l'apprenant. Pour pouvoir faire un matching automatisé entre les deux entrées un enrichissement du modèle LOM est nécessaire. La version actuelle ne prend pas en considération les niveaux de maîtrise d'une langue, alors nous proposons de rajouter des sous catégories à la catégorie 'Langue' que nous avons appelée 'oral', 'lecture' et 'écriture' Ces catégories possèdent les mêmes éléments de la catégorie « Accessibilité » du modèle IMS-LIP.

Cette vue nous amène à proposer un autre filtrage de la langue par rapport au classement sélectionné par l'apprenant au moment de l'interaction avec le système.

Etape4 : la dernière étape consiste à donner un ensemble de briques élémentaires pour un type de média exemple : texte, video...etc.

Après l'utilisation successive de ces quatre filtre, la 5<sup>ème</sup> étape consiste à afficher les meilleures briques élémentaires par l'affectation d'une feuille de style XSL au document résultant.

## 5 Conclusion

Dans ce travail, nous sommes intéressés à la notion d'adaptation dans le domaine de e-Learning, le champ de travail exploré est celui de l'adaptativité de contenu des activités aux objectifs de l'apprenant et l'adaptation de navigation des ressources aux préférences d'apprenant.

Dans notre cas, nous avons choisis d'utiliser le standard IMS –LD pour modéliser le scénario d'apprentissage. Nous avons adopté ce modèle pour sélectionner les activités de l'apprenant qui lui conviennent. Dans le modèle IMS- LD, chaque activité possède un lieu vers un ensemble de ressources ; donc l'apprenant est confronté à une quantité d'information, d'où la nécessité d'un autre niveau d'adaptation, c'est l'adaptation aux préférences. Cette adaptation, nous amène à adopter le LOM, pour l'indexation des objets pédagogiques. Nous avons proposé une extension du LOM qui nous facilite le filtrage des briques selon les préférences de l'apprenant. Par ailleurs, les recherches dans les standards pour modéliser l'utilisateur, nous amènent à adopter le standard IMS–LIP qui définit une structure de données pour pouvoir être importée ou exportée entre les systèmes interopérables.

Pour la gestion d'adaptation, nous nous sommes basés sur la technologie XML et plus particulièrement XSLT. L'adaptation est un domaine très vaste, à court terme, nous pensons réfléchir à évoluer les dimensions d'adaptabilité. Pour ce faire nous envisageons de poursuivre le développement de l'environnement d'apprentissage par traçabilité de déroulement d'un apprenant dans un scénario. Il est également envisageable de prendre en compte l'adaptation dans une situation d'apprentissage collaboratif basé sur le modèle de groupe d'utilisateurs.

## Références

1. Alfabet 2003 <http://alfanet.ia.uned.es/>
2. Alonso M., « Modélisation, adaptation et opérationnalisation de scénarios d'apprentissage » Rapport de Master 2eme année, Université Joseph Fourier, Juin 2005
3. Brusilovsky P., Methods and Techniques of Adaptive Hypermedia, User Modeling and User-Adapted Interaction, (1996)Vol.6, pp. 87-127.
4. Brusilovsky P., Peylo C., Adaptive and Intelligent Web-based Educational Systems, *International Journal of Artificial Intelligence in Education*, (2003)pp.156-169.
5. Delestre N. « METADYNE, un Hypermédia Adaptatif Dynamique pour l'Enseignement », Thèse de l'Université de Rouen, (2000), Soutenue le 20 Janvier 2000.
6. Dieterich. H. Malinowski. U. Kuhme. T schneider-hufschmidt m. State of the Art in Adaptive User Interfaces. M. Schneider-Hufschmidt and U. Malinowski Eds. (1993) pp 13-48. 1993
7. Dufresne Aude article « Conception d'une interface adaptée aux activités de l'éducation à distance – ExploraGraph »2001. [dufresne@com.umontreal.ca](mailto:dufresne@com.umontreal.ca)
8. Faerber R., Caractérisation des situations d'apprentissage en groupe, *revue STICEF*, (2004) Vol. 11, pp.297-331.
9. E. P. Giacomini, "Netuniversité, une plate-forme basée sur IMS-LD pour la conception de cours en ligne dans le cadre du projet CEPIAH, (conception et évaluation des polycopiés interactifs pour l'apprentissage humain », (2005) Thèse de Doctorat de l'université de Compiègne, Nov 2005.
10. Habeib-mammar EDPHA: « un Environnement de Développement et de présentation d'Hyperdocuments Adaptatifs » (2004) Thèse de Ecole Doctorale Informatique et Informatique pour la Société, soutenue le 10 septembre 2004.
11. IMS Global Learning Consortium. IMS Learning Design Specification (2003).
12. Koper R., Tattersall C. (2005), *Learning Design*, The Netherlands, Springer.
13. LOM (2002) Final Draft Standard for Learning Object Metadata, Approved draft, Document IEEE 1484.12.1-2002, 2002. 44p. <http://ltsc.ieee.org/wg12/>
14. Malevergne Emeric Expression des intentions didactiques et réutilisation des ressources pédagogiques Rapport de DEA Environnements Informatiques pour l'Apprentissage Humain et Didactique (2003).
15. Pernin.P, A. Lejeune "Dispositifs d'Apprentissage Instrumentés par les Technologies: vers une ingénierie centrée sur les scénarios", actes du colloque TICE 2004, p.407-414, Compiègne, octobre 2004.

## *Intelligent Systems*

---





## Prédiction de la structure des protéines par l'approche neuronale

Hafida Bouziane<sup>1</sup>, Belhadri Messabih<sup>1</sup>, Abdellah. Chouarfia<sup>1</sup> et Ghania Ghensous<sup>2</sup>

Université des Sciences et de la Technologie d'Oran

USTO, BP 1501 El M'Naouer, Oran ALGERIE

E-mail : <sup>1</sup>{h\_bouziane,messabih,chouarfia} @univ-usto.dz

<sup>2</sup>Khensous.info@gmail.com

**Résumé.** Un grand nombre de protéines séquencées ne trouvent pas dans les banques de données des séquences qui leur sont voisines et qui permettraient de part leur apparentement, de déterminer leur structure et/ou leur famille fonctionnelle par correspondance avec une protéine dont la structure et/ou la fonction est déjà connue.

La Prédiction de la structure protéique est un intérêt majeur. La nature de la Protéine, sa structure, son mode de fonctionnement, son comportement...etc. font l'objet de plusieurs recherches. Il existe en effet, différentes méthodes basées sur des algorithmes de prédiction fournissant des informations sur la structure secondaire à partir de la structure primaire (séquence en acides aminés) sans aucun travail biochimique supplémentaire. Dans ce papier nous proposons une approche neuronale basée sur l'apprentissage par des réseaux de neurones multicouches du type feed-forward tels les réseaux à fonctions radiales (Radial Basic Fonctions) et les perceptrons multicouches (Multi Layer Perceptrons).

**Mots clés :** Bioinformatique, Protéine, Acides aminés, Structures secondaire, Apprentissage, Réseaux de neurones.

### 1 Introduction

La protéine est une macromolécule composée d'acides aminés (les éléments de base) reliés par des liaisons peptidiques [1]. Elle est présente chez les organismes vivants et essentielle à leur fonctionnement. Prédire la structure 3D d'une protéine, c'est approcher la fonction potentielle qu'elle assure dans la cellule. Les structures 2D et 3D sont mieux conservées que les séquences (structures 1D) au cours de l'évolution d'où l'intérêt cognitif majeur. Au niveau tridimensionnel, les protéines possèdent des conformations périodiques très particulières (structures secondaires) qui sont essentiellement de trois types : hélices  $\alpha$ , feuillets  $\beta$  et boucles. Une étape très importante du travail est de prévoir où sont localisés les éléments de structures secondaires sachant que l'information nécessaire au repliement des protéines dans leur conformation biologiquement active est contenue dans la séquence. Il est difficile, voire impossible de prédire la structure complète de la protéine. On propose donc de fragmenter cette séquence en B.P (Blocs Protéiques) appelés « fenêtres d'apprentissage ». L'information est une fenêtre de 13 à 17 acides aminés (résidus) de la séquence requête. L'analyse du signal à travers le réseau neuronal donne alors le type de structure secondaire du résidu central de la fenêtre.

### 2 Prédiction de la structure 2D protéique

La structure secondaire correspond aux formations périodiques qu'adoptent des portions partielles d'une protéine donnée [1]. Ces repliements sont : l'hélice  $\alpha$ , le feuillet  $\beta$  et une structure apériodique (chaîne aléatoire ou coude). Néanmoins, d'autres structures régulières moins répandues existent : l'hélice  $3_{10}$ , les  $\beta$  coudes et les Y coudes. La fonction d'une protéine dépend en grande partie de sa structure. Prédire la structure 2D revient à prédire les éléments conformationnels locaux (hélices  $\alpha$ , feuillets  $\beta$  et structures apériodiques). La structure secondaire est codifiée en une succession d'états par un alphabet de trois lettres : portion en hélice  $\alpha$  (H), feuillet  $\beta$  (E : brin étendu) et en coude (C). Les réseaux de neurones modélisant un type de structure secondaire sont entraînés par des séquences dont la structure est connue.

### 3 Modélisation par un réseau de neurone

Un réseau de neurones artificiel est basé sur un modèle simplifié de neurones. Comme le montre la figure1, il doit pouvoir calculer des valeurs de sorties (s1, s2, s3) en fonction de variables d'entrées (e1, e2, e3).

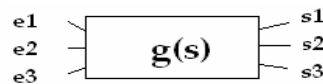


Fig.1. Schéma général d'un réseau de neurones

Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle.

### 3.1 Le perceptron multicouche (Multi Layer Perceptron)

Le perceptron est un des premiers réseaux de neurones, conçu en 1958 par Rosenblatt [2]. Il est linéaire et monocouche. Les sorties des neurones ne peuvent prendre que deux états (-1 et 1 ou 0 et 1). La règle de modification des poids est celle de Widrow-Hoff. En 1969, Papert & Minsky [3] ont démontré ses limites. Le perceptron multicouches est une amélioration du perceptron comprenant une ou plusieurs couches intermédiaires (couches cachées).

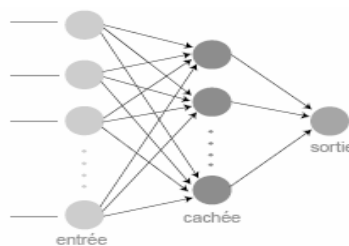


Fig.2. Exemple de perceptron multicouches

Pour modifier les poids, l'algorithme de rétropropagation du gradient est le plus utilisé, la sigmoïde représente la fonction de base de ces réseaux [4].

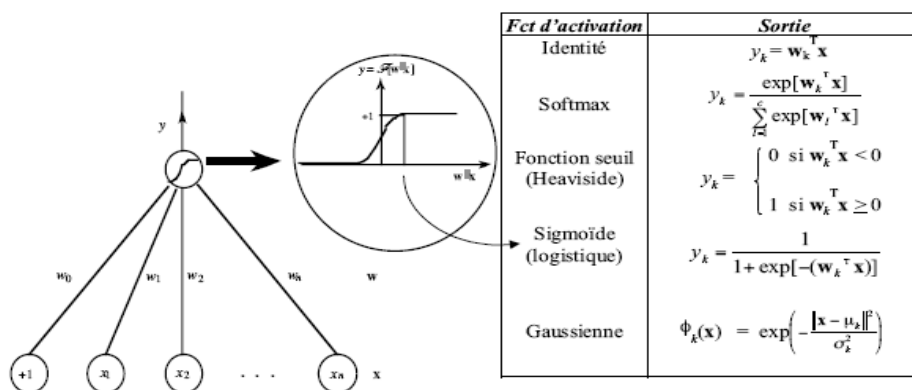


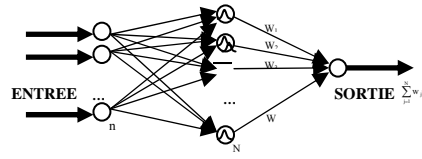
Fig.3. La sigmoïde : fonction d'activation pour les réseaux MLP

Nous présenterons dans la partie expérimentale quelques résultats obtenus. L'utilisation de ce modèle de réseau dans notre recherche est décrite dans notre papier référencé [5].

### 3.2 Réseau à base de fonctions radiales (Radial Basic Functions)

#### 3.2.1 Architecture du réseau

Un réseau RBF est un réseau de neurones de type *feed-forward*, constitué de 3 couches. Chaque neurone d'une couche est connecté à tous les neurones de la couche suivante [6]. La couche d'entrée retransmet les entrées sans distorsion aux neurones de la couche intermédiaire. La couche cachée contient les neurones de type RBF (fonctions de base ou noyaux). La couche de sortie contient les valeurs de sorties du réseau. Chaque vecteur dans l'ensemble d'apprentissage est assigné à un neurone dans la couche cachée, ce vecteur est dit *prototype*. La couche intermédiaire calcule une fonction Gaussienne de la distance entre le vecteur en entrée (vecteur à classifier) et le *prototype*.



**Fig.4.** Réseau RBF avec ses trois couches.

L'équation d'un réseau RBF se résume donc en :

$$\hat{y} = \sum_{i=1}^k \lambda_i e^{\left( \frac{\|x - c_i\|^2}{2\sigma_i^2} \right)} \quad (1)$$

Où  $x$  est le vecteur en entrée,  $\hat{y}$  la sortie scalaire du réseau RBF,  $c_i$  les centres des fonctions Gaussiennes,  $\sigma_i$  leurs largeurs respectives et  $\lambda_i$  les poids du réseau, ils correspondent à l'importance relative de chaque noyau dans la sortie du réseau  $\hat{y}$ .

### 3.2.2 Les paramètres du réseau

Il y a 4 paramètres principaux [6] : Le nombre de neurones dans la couche cachée. La position des centres des gaussiennes de chacun des neurones. La largeur de ces gaussiennes. Le poids des connexions entre les neurones de la couche cachée et le(s) neurone(s) de sortie.

### 3.2.3 Techniques d'apprentissage

L'apprentissage des RBF permet de déterminer les paramètres cités ci-dessus. Il peut se faire selon trois techniques : supervisées, heuristiques et en deux temps.

#### 3.2.3.1 Techniques supervisées

Le principe de ces techniques est de minimiser l'erreur quadratique :

$$E = \sum_n E_n \quad (2)$$

Avec :

$$E_n = \frac{1}{2} \sum_k (\zeta_k^n - y_k(x^n))^2 \quad (3)$$

Où  $\zeta_k^n$  représente la sortie désirée du  $k^{\text{ème}}$  neurone de sortie pour l'exemple  $n$  et  $y_k(x^n)$  sa sortie réelle par rapport à l'entrée  $x$  de l'exemple  $n$ . En utilisant la fonction gaussienne et en considérant les variations suivantes :  $\Delta W_k$  pour les poids des connexions entre la couche intermédiaire et celle de sortie,  $\Delta \mu_j$  pour les centres des prototypes et  $\Delta \sigma_j$  pour les rayons d'influence des fonctions gaussiennes. A chaque pas d'apprentissage, la loi de mise à jour est obtenue en utilisant la descente de gradient sur  $E_n$  :

$$\Delta w_{kj} = \eta_1 (\zeta_k^n - y_k(x^n)) \phi_j(x^n) \quad (4)$$

$$\Delta \mu_j = \eta_2 \phi_j(x^n) \frac{\|x^n - \mu_j\|}{\sigma_j^2} \left( \sum_k (\zeta_k^n - y_k(x^n)) w_{kj} \right) \quad (5)$$

$$\Delta \sigma_j = \eta_3 \phi_j(x^n) \frac{\|x^n - \mu_j\|^2}{\sigma_j^3} \left( \sum_k (\zeta_k^n - y_k(x^n)) w_{kj} \right) \quad (6)$$

Avec  $\eta_1$ ,  $\eta_2$  et  $\eta_3$  représentant les paramètres d'apprentissage.

Cette approche exige un temps d'apprentissage aussi important que celui du MLP.

### 3.2.3.2 Techniques heuristiques

Généralement, le réseau est initialisé sur un centre avec un rayon d'influence ( $\mu_0, \sigma_0$ ). Les centres des prototypes  $\mu_i$  sont créés itérativement au fur et à mesure de la présentation des vecteurs d'apprentissage. La seconde étape consiste à modifier les rayons d'influence et les poids des connexions ( $\sigma_i, w_i$ ) (uniquement les poids entre la couche intermédiaire et la couche de sortie).

#### ○ Algorithme R.C.E. (Restricted Coulomb Energy)

L'algorithme modifie l'architecture du réseau d'une façon itérative durant l'apprentissage. Les neurones intermédiaires ne sont ajoutés que lorsque cela est nécessaire. Un seuil  $\theta$  permet d'ajuster les rayons d'influence.

#### ○ Algorithme D.D.A. (Dynamic Decay Adjustment)

Il est extrait partiellement de l'algorithme RCE [7]. Il consiste à introduire deux seuils  $\theta^-$  et  $\theta^+$  pour réduire les zones de conflit entre prototypes (problème essentiel rencontré dans l'algorithme RCE). Pour assurer la convergence, le réseau doit satisfaire les deux inégalités ci-après pour chaque vecteur  $x$  de classe  $c$  de la base d'apprentissage.

$$\begin{aligned} \exists i : \phi_i^c(x) &\geq \theta^+ \\ \forall k \neq c, \forall j : \phi_j^k(x) &< \theta^- \end{aligned} \quad (7)$$

### 3.2.3.3 Techniques d'apprentissage en deux temps

Les paramètres du RBF sont estimés en deux phases : une première phase détermine les centres et les rayons des fonctions de base à l'aide des vecteurs d'entrée uniquement. La deuxième phase calcule les poids des connexions de la couche cachée vers la couche de sortie (apprentissage supervisé).

#### Première phase (Non Supervisée)

##### ○ Segmentation en k - moyennes des centres (K-Means Algorithm)

Les centres des prototypes et la matrice des variances d'un RBF peuvent être calculés en deux phases. Dans la première phase, l'algorithme détermine le centre du nuage de  $N^j$  points faisant partie de la même classe. Chaque itération comporte deux étapes. Une étape d'affectation de l'ensemble des points d'apprentissage appartenant à la classe  $c$  aux  $k$  centres (de la même classe) en minimisant la variance intragroupe. Le minimum s'obtient en affectant chaque point au centre le plus proche au sens de la distance euclidienne. On obtient ainsi une segmentation de l'espace d'apprentissage  $\chi_c$  en  $k$  groupes disjoints  $\{\chi_j\}_{j=1}^k$  de  $N^j$  points. Puis une étape de minimisation faisant décroître à nouveau la variance intragroupe en calculant le nouveau centre de chaque groupe par la moyenne  $\hat{\mu}_j$  :

$$\mu_j \approx \hat{\mu}_j = \frac{1}{N^j} \sum_{x \in \chi_j} x \quad (8)$$

La convergence de l'algorithme est prononcée à la stabilité des centres des nuages de points. On calcule alors la matrice de covariance de la fonction gaussienne. Celle-ci est estimée par l'expression de la covariance simple ci-dessous :

$$\sigma_j^2 = \hat{\sigma}_j^2 = \frac{1}{N^j} \sum_{x \in \chi_j} (x - \hat{\mu}_j)(x - \hat{\mu}_j)^T \quad (9)$$

##### ○ Méthode E.M. (Expectation Maximisation)

L'algorithme détermine, de manière itérative les paramètres d'un mélange de gaussiennes (par le maximum de vraisemblance). Il permet ainsi d'obtenir les paramètres du réseau en deux étapes : l'étape E calculant la valeur de l'espérance mathématique des données inconnues par rapport aux données connues et l'étape M maximisant le vecteur des paramètres de l'étape E.

#### Deuxième phase (Supervisée)

##### ○ Maximum d'appartenance

Cette technique prend les valeurs les plus importantes des fonctions de base  $\Phi_i(x)$  :

$$\phi_{\max} = \max_{i=1}^N \Phi_i(x) \quad (10)$$

Où  $N$  est le nombre de fonctions de base pour toutes les classes. La sortie du réseau de neurones est alors donnée par :

$$y = \text{classe}(\Phi_{\max}) \quad (11)$$

○ **Algorithme des moindres carrés (L.M.S. : Least Mean Squares)**

Soit à minimiser une fonction de risque empirique ( $R_{\text{emp}}$ ), la détermination des paramètres se fait par descente de gradient. Si la fonction de coût choisie est quadratique à fonctions de base fixées  $\Phi$ , la matrice de pondération  $w$  est obtenue par simple résolution d'un système linéaire. Soit, en effet, un échantillon d'apprentissage :

$$\{(x^n, \zeta^n)\}_{n=1}^N \quad (12)$$

On suppose que le Réseau à Fonctions de base Radiales comporte  $M$  sorties, on cherche les poids  $w$  qui minimisent le risque empirique :

$$R_{\text{emp}} = \sum_{n=1}^N \sum_{i=1}^M (y_i(x^n) - \zeta_i^n)^2 = \sum_{n=1}^N \sum_{i=1}^M \left( \sum_{k=1}^J w_{ik} \phi_k(x^n) - \zeta_i^n \right)^2 \quad (13)$$

Les conditions d'optimalités sont donc obtenues en annulant la dérivée de cette quantité par rapport à  $w_{ik}$ , on obtient alors :

$$\sum_{n=1}^N \sum_{k=1}^J w_{ik} \phi_k(x^n) \phi_k(x^n) = \sum_{n=1}^N \zeta_i^n \phi_k(x^n) \quad (14)$$

Qui peut s'écrire sous la forme matricielle suivante :

$$(\Phi^t \Phi) w^t = \Phi^t \zeta \quad (15)$$

Si la matrice  $\Phi^t \Phi$  est non singulière<sup>1</sup>, la solution optimale pour les poids à fonctions de base fixées, s'écrit :

$$w^t = (\Phi^t \Phi)^{-1} \Phi^t \zeta \Rightarrow w^t = \Phi^{-1} \zeta \quad (16)$$

#### 4 Critère d'évaluation

Plusieurs techniques peuvent être utilisées pour calculer le degré d'efficacité d'un algorithme de prédiction. Le test Q3 donne le pourcentage de résidus pour lesquels la structure secondaire a correctement été prédite. Cet indice représente le taux de succès des prédictions par rapport à l'échantillon [8],[9].

$$Q_3 = \frac{\sum_{i \in SS} a_{ii}}{N} \quad (17)$$

Où  $a_{ii}$  est le nombre de résidus observés dans l'état  $i$  et prédits dans l'état  $i$ ,  $SS$  l'ensemble des états de structure secondaire et  $N$  le nombre total de résidus.

#### 5 Expérimentation

La base d'apprentissage utilisée est celle de Rost et Sander [10] contenant 126 protéines ayant moins de 25% d'identité et ses séquences protéiques sont alignées à l'aide du programme CLUSTAL-W<sup>2</sup>, sachant que l'alignement multiple permet d'augmenter l'efficacité à 10%. Chaque acide aminé est présenté au réseau par un vecteur binaire de taille 21 : 20 acides aminés + un symbole de terminaison. Ces 20 acides aminés sont groupés en 4 classes codifiées comme suit :

H: Alpha Helix (hélices  $\alpha$ ). E: Extended strands (feuillets  $\beta$ ). T: Beta Turns (virages, coudes). C: Random Coils (chaînes aléatoires).

Le réseau RBF est constitué de 3 couches : Entrée, Cachée, et Sortie.

La couche d'entrée contient 17 nœuds, dont chacun correspond à un acide aminé à classifier. La couche cachée est constituée initialement de 4 nœuds, chacun permet de calculer une fonction gaussienne, ils correspondent aux 4 classes (H, E, T, C), cette couche est mise à jour par l'algorithme DDA lors de

<sup>1</sup> Condition respectée pour les fonctions Gaussiennes

<sup>2</sup> Programme informatique pour l'alignement multiple téléchargeable à l'European Biological Institut : <http://ftp.ebi.ac.uk/pub/software/unix/clustalw/>

l'apprentissage. La couche de sortie : correspond aux résultats de la prédiction, l'acide aminé en entrée appartiendra donc à l'une des 4 structures : hélice, feuilles, coude ou chaîne aléatoire.

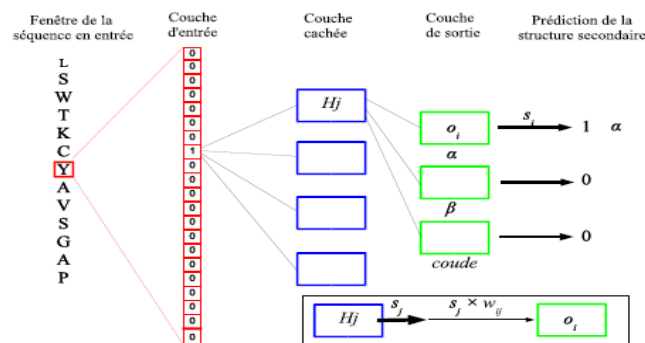


Fig.5. Prédiction de la structure 2D protéique par un réseau de neurones à trois couches

## 6 Résultats et discussions

Comme tout autre RNA, les problèmes majeurs rencontrés lors de l'implémentation de l'algorithme DDA, sont essentiellement l'initialisation des paramètres suivants :

**Le choix des prototypes :** l'implémentation de l'algorithme DDA nécessite une initialisation d'un représentant de chaque classe. Le problème qui se pose est le choix du prototype qui représente au mieux une classe. Généralement, ces centres de classe sont sélectionnés aléatoirement.

**Le choix des largeurs :** le deuxième problème est l'initialisation des largeurs ( l'écart type) des gaussiennes, car le nombre des neurones dans la couche cachée dépend fortement des largeurs initiales attribuées à chaque fonction. Pour palier cet inconvénient, nous proposons des valeurs aléatoires, dans la mesure de la logique ou les structures hélices et feuilletts dominent la conformation secondaire, et qui sont plus significatives quant au fonctionnement et l'évolution de la protéine.

Le test Q3 est évalué au fur et à mesure sur l'ensemble des protéines. Nous avons pu avoir plus de 70% de séquences d'acides aminés correctement prédites. Le Q3 a atteint 60 % pour des séquences de plus de 100 acides aminés. Pour l'approche MLP, L'apprentissage a été lancé dans une boucle infinie pour aboutir à une erreur stable invariante. Le Q3 a pu franchir 80% dans une séquence de moins de 100 acides aminés (résidus). Dans le tableau suivant nous donnons quelques résultats obtenus.

**Table1.** Pour des séquences de 200 à 250 acides aminés, le Q3 a varié de 63% à 60% et a commencé à faiblir pour les protéines dont la teneur en acides aminés augmentait.

Nombre d'acides aminés	752	327	272	150	279	39	29
Test Q3	34,2%	46,8%	51,1%	54,7%	61,6%	74,4%	82,8%

## 7 Conclusion

Dans cet article nous avons montré que les réseaux de neurones restent toujours les éléments de base des méthodes de prédiction statistiques. Par leur caractère de systèmes paramétrés reliant des variables d'entrée à des variables de sortie, ils restent toujours compétitifs. L'apprentissage des réseaux RBF dépend fortement de ses quatre paramètres à savoir : le nombre des gaussiennes, leurs centres, leurs largeurs et les poids synaptiques. L'ajustement de ces paramètres permet de converger après un nombre d'itération plus au moins réduit comparé aux réseaux MLP. En pratique, l'utilisation des méthodes neuronales pose certaines difficultés. La principale étant l'optimisation de la phase d'apprentissage. Le choix de l'architecture adéquate se fait par essais successifs. La deuxième difficulté est liée à l'initialisation qui fait converger le réseau vers l'erreur locale (MLP). Une hybridation avec les algorithmes génétiques pourrait aider à converger vers l'erreur globale. Néanmoins, l'optimisation des paramètres donnera sûrement un meilleur taux de reconnaissance.

## **Références**

1. P.C Turner, A.G Mc Lennan, A.D Bates & M.R.H White, « Biologie Moléculaire », (2006).
2. B.Ramal, P.E.Higgins, C.A. «A cursive recognition system based on human reading models, Machine vision Application », V.8.224-231, (1995).
3. M.Minsky, S.Papert, « Perceptron », MIT press, (1969).
4. Gérard Dreyfus, « Reconnaissance de formes par réseau de neurones», Ecole Polytechnique, (2002).
5. H.Bouziane, B.Messabih, A. Chouarfia « Prédiction de la structure 2D des protéines par les réseaux de neurones », International Business Information Management Association IBIMA'08, (2008).
6. <http://info.fundp.ac.be/~hvanpete>.
7. Michael R. Berthold et Jay Diamond, « Boosting the Performance of RBF Networks with Dynamic Decay Adjustment », Advances in Neural Information Processing, (1995).
8. Laurent BLIN, « Prédiction de la structure Secondaire des protéines par apprentissage automatique symbolique numérique », IFSIC, (1996).
9. Yann Guermeur, «Combinaison de Classifieurs Statistiques, Application à la Prédiction de la Structure Secondaire des Protéines », thèse Doctorat, Université Paris 6, (1997).
- 10.B.Rost et C.Sander, « Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks », PNAS-Proceedings of the National Academy of Sciences of the USA, (1993).

## Vers un système d'aide à la décision basé sur les connaissances cruciales

Mohamed TURKI<sup>1</sup>, Faïez GARGOURI<sup>2</sup>

Laboratoire MIRACL

<sup>1</sup> Institut Supérieur des Etudes Technologiques - Sfax Tunisie, BP : 88A Elbustan ; Sfax 3099-  
Tél. : 00216 74431425 ; Fax : 00216 74431386

<sup>2</sup> Institut Supérieur d'Informatique et du Multimédia-Sfax Tunisie,  
Pôle technologique de Sfax, Route de Tunis Km 10 B.P. 242- SFAX 3021-  
Tél. : 00216 74862233 ; Fax : 00216 74862432  
mohamed.turki@isetsf.rnu.tn / faiez.gargouri@fsegs.rnu.tn

**Résumé :** Dans le contexte de l'entreprise étendue, les employés sont confrontés à de nouvelles situations qui accroissent leurs initiatives et leurs responsabilités. Ils ont besoin donc d'accéder à des connaissances individuelles et collectives largement distribuées, pouvant les aider à prendre des décisions dans leurs tâches professionnelles. En effet, la gestion des connaissances peut répondre à ce besoin en proposant des méthodes, outils et techniques adéquats, permettant d'acquérir, préserver et partager les connaissances, dans un but d'opérationnalisation. Par ailleurs, nous proposons dans ce papier une démarche pour la construction d'un système d'aide à la décision basé sur les connaissances cruciales d'une entreprise.

**Mots-clés :** Gestion des connaissances, Système d'aide à la décision, Connaissances cruciales, Ontologie, Fouille de données.

### 1 Introduction

Depuis quelques années, la concurrence se fait rude et le marché est de plus en plus difficile, les entreprises sont influencées conjointement par la mondialisation des marchés, la libéralisation de l'économie et l'impact des technologies de l'information et communication, ce qui engendre des transformations structurelles rapides et un besoin d'accélération des processus de décision. En effet, pour survivre, les entreprises doivent sans cesse améliorer leurs performances; innover sur tous les plans; réduire les cycles et les coûts de conception, de production, de mise sur le marché de leurs produits et services; et accroître leurs réactivités. Cela peut être réalisé par l'adoption d'un processus de gestion de connaissances.

En effet, la gestion des connaissances (ou Knowledge Management) est apparue au début des années 90 et a occupé une place importante dans les organisations. Ce domaine peut être abordé sous des aspects différents selon les disciplines scientifiques : sciences de gestion et informatique. D'une part, dans l'entreprise, les gestionnaires se focalisent sur la création des connaissances et l'apprentissage organisationnel, ce qui renvoie à une étude de l'aspect humain. D'autre part, les spécialistes en informatique, dans le domaine de l'ingénierie des connaissances et système d'information, abordent la problématique de la gestion des connaissances en proposant des outils et des techniques adéquats, permettant d'acquérir et de préserver les connaissances, dans le but de construire une mémoire d'entreprise (ME). Une ME a été définie par [1] comme étant la matérialisation explicite, désincarnée et persistante des connaissances et informations cruciales d'une organisation en vue de faciliter leur accès, partage, réutilisation par les membres de l'organisation, dans leurs tâches individuelles ou collectives.

Cet article est organisé comme suit. Dans la deuxième section, nous présentons le contexte et la problématique traités dans cet article. Dans la troisième section, nous abordons la convergence entre les systèmes d'aide à la décision et la gestion des connaissances. Nous décrivons dans la quatrième section, notre démarche de construction du système d'aide à la décision basé sur les connaissances cruciales. Nous concluons cet article en donnant les perspectives de ce travail de recherche.



## 2 Contexte et problématique

Dans le contexte de l'entreprise étendue, les acteurs sont confrontés à de nouvelles situations qui accroissent leurs initiatives et leurs responsabilités. Ils se transforment ainsi en des acteurs décideurs quelque soient leurs rôles et leurs positions hiérarchiques. Il est donc nécessaire qu'ils accèdent au travers de leurs postes de travail informatisés à des connaissances individuelles et collectives largement distribuées dans le temps et l'espace planétaire de leurs organisations.

Ainsi, le problème que peuvent rencontrer ces acteurs c'est l'insuffisance de connaissances qui peuvent les aider à prendre des décisions. Cette carence de connaissances peut être due d'une part, à des connaissances explicites ou tacites non capitalisées à cause de départ, dispersion d'employés de corps de métiers différents, ou départ d'un sous-traitant. D'autre part, elle peut être due à l'absence de connaissances tacites qui sont incarnées dans les cerveaux de leurs détenteurs, tels qu'habileté ou secret de métier. Pour cela la capitalisation des connaissances s'avère indispensable, elle peut être traitée au plan stratégique et décisionnel dans une organisation ou bien greffée à des activités professionnelles en cours d'opérationnalisation. Cette capitalisation peut faire l'objet de la construction d'une mémoire d'entreprise.

Cependant, d'après [1][5][15], la construction d'une mémoire d'entreprise, ainsi que le développement d'un système d'aide à la décision tel que «Knowledge Warehouse », peuvent coûter cher à l'entreprise et nécessite beaucoup d'investissement dans la mesure où il faut acquérir toutes les connaissances auprès de leurs détenteurs, les préserver, les actualiser et les rendre accessibles par les membres concernés.

Afin de réduire le coût de développement de ces systèmes, ne capitaliser que les connaissances cruciales [13] ou potentiellement cruciales [5] dans un processus d'aide à la décision s'avère une solution intéressante. Ces connaissances doivent être identifiées, analysées puis évaluées avant d'être exploitées.

Dans ses travaux de recherche, Saad [6] s'est inspirée des trois axes du triangle systémique de Le Moigne [9] afin de construire une famille de critères pour la qualification des connaissances. Elle a considéré la connaissance comme un système complexe qui peut être analysé selon les trois pôles suivants:

- *Le pôle fonctionnel* consiste à déterminer le degré de contribution d'une connaissance aux objectifs de l'entreprise.
- *Le pôle ontologique* consiste à déterminer des critères liés aux caractéristiques de la connaissance, afin d'étudier sa vulnérabilité.
- *Le pôle génétique* consiste à prédire la durée d'usage de la connaissance dans l'entreprise, selon ses objectifs à moyen et à long terme.

En effet, Saad a proposé dans ses travaux de recherche, quinze critères d'évaluation de connaissances [8]. Cependant, elle n'a pas pris en considération certains critères, tels que : l'incertitude, l'imprécision [7], et l'aspect sémantique des connaissances.

Néanmoins, l'imprécision et l'incertitude peuvent être mises en valeur par la logique floue. Concernant la sémantique des connaissances (généralisation/spécialisation, agrégation, composition, synonymie, homonymie, équivalence, etc.), elle peut être traitée par l'utilisation des ontologies. En effet, Gruber [4] a défini l'ontologie comme étant une spécification explicite d'une conceptualisation.

Le premier volet de la problématique étant de concevoir et implémenter un Système d'Aide à la Décision Basé sur les Connaissances cruciales (SADBC). Ce système sera utilisé par les acteurs décideurs au cours de leurs activités opérationnelles. D'une part, il doit être capable d'évaluer les connaissances acquises, selon plusieurs critères ou points de vue. Et ce, afin de mesurer le degré de contribution d'une connaissance à la réalisation d'un objectif au cours d'un processus de décision, tenir compte du degré d'incertitude et imprécision de ces connaissances, leur durée de validité, leurs sémantiques, leurs dynamiques et leurs évolutions. D'autre part, le système assiste les acteurs lors de choix d'actions (solutions) stables ou évolutives, en se reposant sur une évaluation "multicritères" de ces actions. Cette évaluation doit être basée sur les connaissances cruciales déjà identifiées, et elle aura comme résultat, des règles de décisions pertinentes qui seront générées par le système.

Néanmoins, ce même système peut être exploité afin d'évaluer la pertinence des connaissances générées (modèles extraits) par les techniques de fouille de données. Cette évaluation se base sur les connaissances capitalisées des experts. En effet, ces connaissances seront prises en considération lors de la phase d'évaluation.

En effet, dans un processus de fouille de données classique, seules les données opérationnelles sont prises en compte dans les phases de pré-traitement (préparation des jeux de données spécifiques) et post-traitement (analyse des modèles de connaissances extraites par l'étape de fouille standard) [11]. Par conséquent, ces modèles risquent d'être triviaux, non pertinents et de ne pas répondre aux attentes des

analystes. Pour cela, l'intégration des connaissances des experts dans la phase de post-traitement de la fouille de données s'avère une solution intéressante.

Généralement, l'intégration des connaissances des experts dans un processus de fouille de données représente un problème essentiel et constitue un défi pour la recherche actuelle dans ce domaine. En effet, certains travaux de recherche proposent des solutions partielles pour cette problématique (Engel 1996, Morik 2000, Euler & Scholz 2004, Bernstein 2005).

Dans ses travaux de recherche [11][12], Brisson a prouvé à travers la mise en œuvre de l'approche KEOPS qui est basée sur un système d'information dirigé par ontologies, la valeur ajoutée de l'intégration des connaissances d'experts tout au long du processus de fouille de données. Son objectif étant alors, la confrontation des motifs extraits par les techniques de fouille de données aux connaissances d'experts acquises sur le domaine. Il envisage ultérieurement, adapter son approche à la fouille de données temporelles. Conjointement, on peut prévoir une évaluation multicritères des connaissances d'experts qui seront utilisées lors de cette confrontation.

Le deuxième volet de notre problématique consiste alors, à utiliser notre système d'aide à la décision, pour évaluer la pertinence des modèles extraits par les techniques de fouille de données, en se basant sur les connaissances des experts. Le but étant de sélectionner parmi les modèles fiables, les plus intéressants et les plus utiles. Les connaissances ainsi retenues seront alimentées dans notre système, afin d'être utilisées dans un processus d'aide à la décision.

### **3 Convergence entre les systèmes d'aide à la décision et la gestion des connaissances**

Dans cette section, nous abordons en premier lieu la position des systèmes d'aide à la décision dans la gestion des connaissances, par la suite nous présentons quelques travaux de recherche relatifs aux systèmes d'aide à la décision basés sur les connaissances.

#### **3.1 Position des systèmes d'aide à la décision dans la gestion des connaissances**

La gestion des connaissances représente aujourd'hui une nouvelle tendance et diligence des entreprises dans le développement de la gestion moderne. En effet, elle surmonte les défauts de la gestion d'information traditionnelle en accentuant la gestion sur la capacité d'apprentissage collaborative ainsi que la créativité. De ce fait, plusieurs compagnies créent une fonction de *responsable de connaissances* (Chief Knowledge Officer : CKO) qui assure, d'une part, la classification des connaissances (explicites et implicites), la conception d'un environnement opérationnel basé sur les connaissances et la gestion de la procédure reliée à ce métier. D'autre part, ce responsable s'occupe de la mise en place d'un support technique pour la gestion des connaissances et la création d'un environnement de communication.

D'après [10], la barrière d'implémentation de la gestion des connaissances implique aussi bien un mécanisme organisationnel et technologique. Le mécanisme organisationnel concerne le droit des organisations, ainsi que la relation entre leurs contrats. Le mécanisme technologique s'intéresse aux méthodes et outils à utiliser pour faire face aux problèmes non-structurés et semi-structurés produits dans la course de la gestion de connaissances.

D'après [10], les systèmes d'aide à la décision assistent les décideurs qui traitent des problèmes semi-structurés. Le but fondamental des ces systèmes étant alors l'amélioration de la qualité de décision à travers l'unification des ressources d'intelligence individuelles et la faculté de traitement informatique.

En effet, la gestion des connaissances dans les entreprises peut tirer profit des avantages des systèmes d'aide à la décision comme; réaliser plusieurs calculs à moindre temps et coût, aider les membres d'un groupe de travail à réduire les problèmes de coordination et de communication, améliorer la productivité, choisir dans un processus de décision la bonne information ou connaissance stockée, évaluer plusieurs alternatives, analyser rapidement le risque et les opinions des experts, partager et découvrir des connaissances, etc.

### 3.2 Travaux relatifs aux systèmes d'aide à la décision basés sur les connaissances

Plusieurs modèles de systèmes d'aide à la décision ont été proposés dans la littérature.

Le modèle proposé par [14] a été conçu pour surmonter les difficultés lors du choix d'un outil multimédia destiné à l'enseignement et l'apprentissage. En effet, ce problème est caractérisé par des connaissances et contraintes incertaines, dynamiques, explicites et implicites, et il fait intervenir plusieurs acteurs. La solution proposée gère et supporte six phases qui sont : planification, intelligence, conception, choix, implémentation et évaluation.

Un autre modèle a été proposé par [10] comporte trois niveaux. Le premier niveau englobe les outils informatiques qui supportent le système d'information d'une entreprise tels que : ERP, CRM, etc. Le deuxième niveau comporte les bases de données et les bases de connaissances multidimensionnelles. Le troisième niveau, contient le système d'aide à la décision basé sur la gestion des connaissances.

Une architecture a été proposée également par [3] et visant à aider les utilisateurs pour la prise de décision dans une opération de commerce électronique. Cette architecture comporte un modèle utilisateur, un modèle de produit, des outils de décisions et un système de conseil qui représente le cœur de l'architecture.

Dans [2], une méthode d'intégration de composants multiples dans un système d'aide à la décision distribué (NED-2) pour la gestion d'un écosystème dans une forêt. Il s'agit d'une approche multi-agents avec une architecture blackboard qui supporte l'expansion du système d'aide à la décision par des nouvelles composantes, parmi lesquelles, on cite, un système d'inventaire sophistiqué, un modèle de faune et flore, des systèmes basés sur les connaissances pour l'analyse de satisfaction et un système de génération de rapports. Dans cette approche une ontologie de domaine a été développée et implémentée dans le modèle de données interne du système.

Finalement dans [6][8] un système d'aide à la décision pour l'identification et l'évaluation des connaissances cruciales a été proposé. L'architecture fonctionnelle de ce système comporte essentiellement quatre composantes qui sont : une interface graphique, une base de modèle, une base de données et une base de connaissances avec un moteur d'inférence.

## 4 Notre démarche de construction du SADBC.

A la lumière des travaux de recherche précédemment cités, on va essayer dans ce qui suit, de proposer une démarche de construction d'un système d'aide à la décision basé sur les connaissances cruciales. Cette démarche comporte essentiellement cinq phases qui sont : planification, spécification, conception, implémentation et évaluation. Elle vise à soutenir des décisions dans des contextes de résolution de problèmes ou choix complexes. On s'est inspiré dans notre proposition, des méthodes d'aide à la décision multicritères présentées dans [16], et du modèle des phases de décision proposé par [17] qui englobe cinq étapes pour la résolution de problèmes.

**Phase de planification :** Après avoir précisé le domaine, le champ d'application et le problème à résoudre, il faudrait définir la politique à suivre pour la mise en place du système d'aide à la décision basé sur les connaissances cruciales.

**Phase de spécification :** Il s'agit de spécifier le problème à résoudre, dresser la liste des actions ou solutions potentielles et identifier les critères qui orienteront les acteurs décideurs. En suite, il faut capitaliser les connaissances cruciales qui sont jugées comme nécessaires lors d'un processus de décision.

**Phase de conception :** En se basant sur ce qui a été spécifié dans la phase précédente, la conception du SADBC peut être abordée. En effet ce système peut contenir plusieurs composantes parmi lesquelles on cite :

- *Les sources de connaissances* tels que, base de données opérationnelles, documents, experts, connaissances générées par un système de fouille de données, internet, etc.
- *Une ontologie de domaine et une ontologie d'application* qui spécifient les concepts et les relations entre ces concepts afin de modéliser la sémantique des connaissances. Ces ontologies peuvent être modélisées puis implémentées dans les modèles de données et de connaissances internes du système.
- *Modèle utilisateur* qui englobe les utilisateurs du SADBC tels que, les experts, le responsable de connaissances, l'analyste et les acteurs décideurs.
- *Une interface utilisateur* qui représente le moyen pour gérer l'interaction entre les acteurs et le système.

- Une base de données qui représente un référentiel de données et des paramètres nécessaires à l'exécution des algorithmes.
- Une base de connaissance qui comporte un référentiel de connaissances sous forme de base de règles de décision et base de faits.
- Une base de modèles qui regroupe les algorithmes nécessaires à l'implémentation du système, tels que les algorithmes d'évaluation de connaissances selon plusieurs critères (fonctionnel, ontologique, génétique, imprécision, incertitude, etc.); les algorithmes d'évaluation des modèles extraits par le processus de fouille de données; les algorithmes de gestion des profils utilisateurs; et les algorithmes d'induction des règles de décisions pour la qualification des connaissances repérées et l'évaluation des modèles extraits.

L'architecture proposée (Fig.1.), représente une première version qui a été inspirée de l'architecture fonctionnelle du Système d'aide à la décision proposée par [8] et enrichie avec d'autres éléments qu'on propose.

**Phase d'implémentation :** Cette étape concerne l'implémentation du SADBC. On devrait définir l'architecture physique du système, et choisir l'environnement de développement tels que : outils, langages, SGBD, moteur d'inférence, etc.

**Phase d'évaluation :** Finalement, il faudrait, valider le SADBC à travers une expérimentation puis le mettre au service du développement.

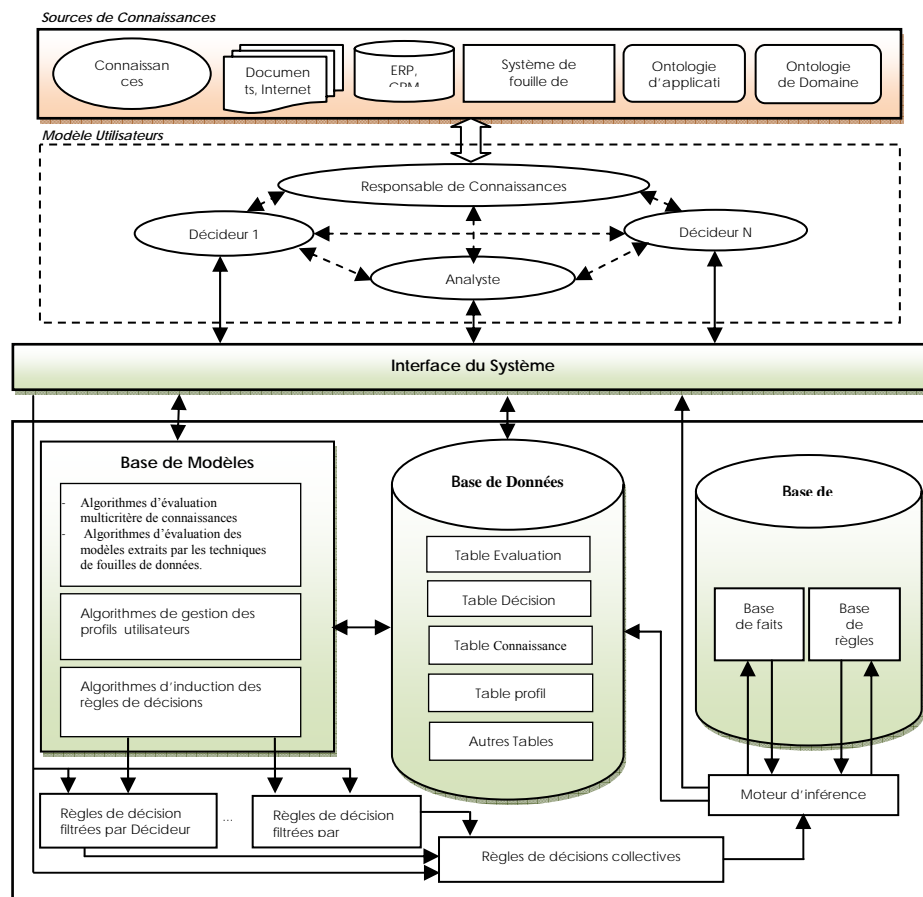


Fig.1. Architecture du SADBC

## 5 Conclusion et perspectives

Dans cet article, nous avons introduit la gestion des connaissances ainsi que les objectifs tirés par son adoption dans une organisation. Nous avons précisé que l'aide à la décision est l'un des objectifs majeurs de ce domaine.

De ce fait, nous avons proposé une démarche de construction d'un système d'aide à la décision basé sur les connaissances cruciales. Notre système assure une évaluation "multicritères" des connaissances acquises et des modèles extraits par un système de fouille de données. Les connaissances retenues par ce système seront utilisées dans un processus d'aide à la décision.

Nos travaux actuels de recherche, portent principalement sur les points suivants : i) Identifier le domaine, le champ d'application ainsi que la politique à suivre pour la mise en place du SADBC. ii) Identifier les moyens à mettre en place pour capitaliser les connaissances cruciales contenues dans des sources différentes. iii) Concevoir le SADBC. iv) Implémenter et mettre en place le système. v) Valider le système à travers une expérimentation puis le mettre au service du développement.

## Références

1. Dieng R., Corby O., Gandon F., Giboin A., Golebiowska J., Matta N. & Ribi re M. «M thodes et outils pour la gestion des connaissances» Dunod 2 me  dition 2001.
2. Donald Nute, Walter D. Potter, Zhiyuan Cheng, Mayukh Dass, Astrid Glende, Frederick Maierv, Cy Routh, Hajime Uchiyama, Jin Wang, Sarah Witzig, Mark Twery, Peter Knopp, Scott Thomasma, H. Michael Rauscher «A method for integrating multiple components in a decision support system». Elsevier 2005.
3. Faisal Al-Qaed, Alistair Sutcliffe «Adaptive Decision Support System (ADSS) for B2C ECommerce» ACM - ICEC'2006.
4. Gruber, T, R.Toward. « Principles for the Design of ontology Used for Knowledge Sharing» Stanford Knowledge Systems Laboratory, 1993.
5. In s Saad, C. Rosenthal-Sabroux et Michel Grundstein : «Une  tude approfondie pour le choix des connaissances   capitaliser en amont de la construction d'une m moire d'entreprise». 2005
6. In s Saad, C. Rosenthal-Sabroux & M. Grundstein, «Improving the Decision Making Process in The Design project by Capitalizing on Company's Crucial Knowledge», Group Decision and Negotiation, Kluwer, volume 14, num ro 2, pp. 131-145, 2005.
7. In s Saad, C. Rosenthal-Sabroux & Salem Chakhar «K-DSS: A Decision Support System for Identifying and Evaluating Crucial Knowledge»-(IPMU) 2006-JULY 2-7, Paris, France.
8. In s Saad, & Salem Chakhar «A Decision Support for Identifying Crucial Knowledge Requiring Capitalizing Operation»-Elsevier-2007.
9. Le Moigne J.L. (1977). «La th orie du syst me g n rale, th orie de la mod lisation» Paris: P.U.F.
10. Liping Sui «Decision Support Systems Based on Knowledge Management» IEEE 2005.
11. L. Brisson, M. Collard, Nicolas Pasquier «Ontologie et base de connaissances pour le pr -traitement et post-traitement en fouille de donn es» EGC 2006.
12. L. Brisson, M. Collard, «Une exp rience d'int gration des connaissances dans un processus de fouille de donn es» ISRN I3SRR-2007-07-FR.
13. Michel Grundstein,Camille Rosenthal-Sabroux «Une aide   la d cision pour le rep rage des connaissances potentiellement cruciales dans un projet de conception-Application du cadre directeur GAMETH» Gestion dynamique des connaissances industrielle. Paris Herm s Lavoisier 2004.
14. Mohammed N. A. Abdelhakim, Shervin Shirmohammadi «A Web-Based Group Decision Support System for the Selection and Evaluation of Educational Multimedia» ACM - EMME'07, September 28, 2007.
15. Nemati, Steiger, Iyer, Herschel «Knowledge Warehouse: An Architectural integration of Knowledge Management, Decision support, Artificial Intelligence and Data Warehousing»; Decision Support Systems 33, Elsevier, (2002), 143-161.
16. P. Njanda «Note M thodologique sur l' valuation Multicrit re de logiciels SIG» 09-2006.
17. Turban Efraim and Aronson, Jay E. (2001). «Decision Support Systems and Intelligent systems» (sixth Edition). Prentice Hall, NJ, USA.

## Vers la Conception d'un Modèle de Classification par Automates d'Arbres

Souad Taleb Zouggar et Baghdad Atmani

Département Informatique, Faculté des Sciences, Université d'Oran  
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie  
[souad.taleb@gmail.com](mailto:souad.taleb@gmail.com), [atmani.baghdad@univ-oran.dz](mailto:atmani.baghdad@univ-oran.dz)

**Résumé.** L'élaboration des modèles de classification à partir d'exemples est un champ de recherches ancien et toujours actif dans la communauté apprentissage automatique. Plusieurs algorithmes dédiés à cette tâche ont été mis au point. Nous nous intéressons particulièrement aux techniques symboliques à base d'arbres de décision. Dans cet article, nous présentons notre approche, en cours d'étude, pour la génération d'arbres de décision en nous attachant à un formalisme ayant fait ses preuves dans plusieurs domaines de l'informatique : les automates d'arbres. Notre contribution dans ce domaine concerne l'optimisation ou le post-élagage des modèles de classification à base d'arbres de décision. Nous nous sommes fixés comme objectifs l'expérimentation des algorithmes de minimalisation d'automates d'arbres pour simplifier les modèles de classification générés.

**Mots-clés :** Classification, Apprentissage automatique, Arbres de décision, Post-élagage, Automates d'arbres.

### 1 Introduction

Le problème de classification est un problème d'ECD (Extraction de connaissances à partir de données). Plusieurs méthodes ont été mises au point pour résoudre ce problème [21,24,25]; parmi lesquelles, on retrouve les méthodes statistiques, à base de réseaux de neurones [1], à base d'arbres de décision [2], etc.

Les méthodes à base d'arbres de décision présentent un grand intérêt pour les chercheurs car ce sont des méthodes intelligibles; elles représentent graphiquement un ensemble de règles, dites règles de classification, qui sont facilement interprétables.

Notre contribution s'inscrit dans le domaine de l'apprentissage automatique à partir de données et utilise comme base de travail les méthodes de classification supervisées à base d'arbres de décision. L'objectif est la conception et l'expérimentation de technique de génération et d'optimisation des arbres de décision par une méthode formelle de modélisation. Pour cela, nous allons expérimenter la théorie des automates et en particulier celle des automates d'arbres ; qui sont des modèles de calcul à états finis généralisant les automates classiques à la reconnaissance de termes.

La finalité de ce travail est de simplifier le modèle d'arbre traduit dans le formalisme d'automate d'arbre en utilisant les algorithmes de simplification d'automates d'arbres existants.

Cet article est structuré comme suit. La section 2 est consacrée aux automates d'arbres, et en particulier à la présentation d'une synthèse des différents travaux de recherches à base de cet outil formel dans le domaine de l'apprentissage automatique. Dans la section 3 nous illustrons à travers un exemple notre contribution à base d'automates d'arbres. Enfin, nous concluons par la section 4.

### 2 Etat de l'Art sur les Automates d'Arbres

Les automates d'arbres [10] ont été employés à l'origine pour la vérification de circuits et ont été ensuite souvent utilisés dans le cadre de l'interprétation abstraite à partir de contraintes, dans le contexte de problèmes de réécriture, de preuves automatiques de théorèmes ou encore de vérification de programmes. Ils ont également été utilisés pour l'inférence de programmes logiques. Leur utilisation s'est récemment étendue dans des applications diverses :

Dans le domaine de traitement de données structurées en arbres, telles que les données au format XML, les travaux de Dal Zilio et Lugiez [28] consistaient à la vérification de la pertinence d'un modèle réalisé à partir d'automates d'arbres pour la compilation de langage existants pour l'interrogation de documents XML. Les mêmes auteurs [29] ont aboutis vers une nouvelle classe d'automates : les sheaves automata et une logique dérivée de la logique des ambients dédiés à l'interrogation de ces documents. Dans le même contexte Dal Zilio et Acciai [31] utilisent une extension de la classe d'automates d'arbres proposée dans [30,32] et proposent des opérateurs de filtrage pour l'interrogation de tels documents.

Dans le domaine de la validation de documents XML, Bouchou et al. [5] proposent d'utiliser les automates d'arbres pour la vérification des contraintes de clés des documents suite à une mise à jour et dans un autre travail des mêmes auteurs [6] les automates sont utilisés pour tester de façon incrémentale, en vérifiant seulement la partie du document concernée par les mises à jour, la validité d'un document.

Les protocoles cryptographiques sont des spécifications de suites de messages entre agents, utilisant des moyens de chiffrement (entre autres), et cherchant à assurer des besoins en confidentialité, authentification, ou d'autres propriétés. Leur vérification par les outils classiques s'est avérée coûteuse ou non satisfaisantes. David Monniaux [18] a proposé une nouvelle approche de vérification basée sur l'interprétation abstraite et l'utilisation de langages réguliers d'arbres, Genet et Klay [12] proposent une nouvelle technique basée sur des systèmes de réécriture et sur des automates d'arbres. Toujours dans le domaine de vérification, Genet et son équipe [14] réalisent la preuve d'une manière automatisée sur un modèle de réécriture de termes du protocole utilisant TIMBUK [13]; un outil de vérification en utilisant l'interprétation abstraite sur des domaines d'automates d'arbres.

Dans le domaine de l'apprentissage automatique, le travail de Carrasco et al. [9] consiste à généraliser un ancien algorithme pour l'identification de langages réguliers à partir d'exemples stochastiques au cas de langages d'arbre. Il sera décrit aussi une méthode de calculer efficacement l'entropie relative entre la grammaire cible et celle inférée, utile pour l'évaluation de l'inférence. Habrard et son équipe [16] étudient le problème de l'apprentissage d'une distribution statistique de données dans une base de données relationnelle en proposant une méthode capable de prendre en compte la structure de la base de données et ne nécessite pas de transformations de données qui peuvent engendrer une perte d'informations essentielles. Le travail d'Habrard [17] dans le cadre de l'inférence grammaticale s'intéresse principalement à deux problématiques : le traitement de données bruitées ou non pertinentes, et l'extraction de connaissances à partir de données arborescentes. Le travail de Tommasi [26] aborde la question sous l'angle de la réalisation automatique de programmes d'annotation d'arbres, permettant de dériver des procédures de transformation ou d'exécution de requêtes.

Dans un contexte applicatif, plusieurs systèmes à base d'automates d'arbres ont été mis au point, à titre d'exemple, le système GIFT (Grammatical Inference for Terms) [3,4] apprend un automate d'arbres à partir d'un ensemble de termes, qui par la suite sera traduit en un programme logique, l'algorithme est appliqué à des données structurées, et un système de typage est aussi inféré pour éviter les situations incorrectes. CPV (Cryptographic Protocol Verification) [15] est un logiciel de vérification automatique de propriétés de confidentialité de protocoles cryptographiques fondé sur des techniques d'automates d'arbres.

Le système Squirrel [8] utilise l'algorithme d'apprentissage de langages d'arbres d'arité non bornée qui dérive de RPNI [19]. Le travail a été formulé pour les séquences mais qui repose sur des principes que l'on retrouve à la fois dans les automates de mots et dans les automates d'arbres.

Enfin, pour terminer ce bref historique sur les automates d'arbres nous soulignons qu'il existe des outils destinés à la manipulation d'automates d'arbres comme par exemple TIMBUK [13], etc.

### 3 Vers la Conception d'un Modèle d'Automates d'Arbres pour la Classification

#### 3.1 Définition d'un Automate d'Arbres [17]

Un automate d'arbres fini  $A$  est un quadruplet  $A = (Q, \Sigma, \delta, Q_f)$  où :

1.  $\Sigma$  Signature définie par le quadruplet  $(\tau, V, \text{arité}, \sigma)$  :
    - $\tau$  : Ensemble fini dont les éléments sont des types,
    - $V$  : Alphabet dont les éléments sont des symboles fonctionnels,
    - $\text{Arite} : V \rightarrow N$  arité ( $f$ ) appelée l'arité de  $f$ ,
    - $\sigma : V \rightarrow \tau$   $\sigma(f)$  est appelée le type de  $f$ ,
  2.  $Q$  Ensemble fini d'états,
  3.  $\delta : V \times Q^* \rightarrow Q$  Fonction de transition (voir figure 1),
  4.  $Q_f \subseteq Q$  Ensemble d'états finaux ou d'acceptation.
- Pour les automates ascendants, les règles de transition sont de la forme :  $\delta(f, q_1, \dots, q_s) = q$  pour un symbole  $f \in V$  dont le nombre d'arguments ( $\text{arité} = s > 0$ ). Ces règles peuvent aussi s'écrire :  $f(q_1, q_2, \dots, q_s) \rightarrow q$ .
- Un arbre  $t$  est accepté par l'automate ascendant  $A$  ssi  $\exists q \in Q_f ; \delta'(t) \rightarrow q$ ,  $\delta'$  correspond à plusieurs applications de  $\delta$ .
- Pour les automates descendants, le quadruplet est  $A = (Q, \Sigma, \delta, Q_f)$  où  $Q_f \subseteq Q$  est l'ensemble des états initiaux ; l'ensemble des états finaux est remplacé par l'ensemble des états initiaux, une règle de transition est de la forme :

$$q \rightarrow f(q_1, q_2, \dots, q_s).$$

Un arbre  $t$  est accepté par l'automate descendant  $A$  ssi  $\exists q \in Q_f ; q \rightarrow \delta^*(t)$ .

– Une règles- $\varepsilon$  est une règle reliant deux états, elle est de la forme :  $q_i \rightarrow q_j ; q_i, q_j \in Q$ .

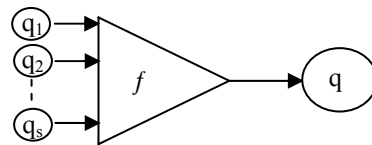


Fig. 1. Représentation graphique d'une transition.

### 3.2 Contribution à Travers un Exemple

Dans cette section, il s'agit de présenter un exemple et de réaliser pas à pas la traduction de l'arbre généré par ID3 dans le formalisme d'automate d'arbres. L'exemple se compose de 04 variables descriptives, une classe « Jouer\_Tennis » et 14 individus.

Table 1. Echantillon d'apprentissage [20]

Ciel	Température	Humidité	Vent	Jouer
Ensoleillé	Chaude	Elevée	Faible	Non
Ensoleillé	Chaude	Elevée	Fort	Non
Couvert	Chaude	Elevée	Faible	Oui
Pluie	Tiède	Elevée	Faible	Oui
Pluie	Fraîche	Normale	Faible	Oui
Pluie	Fraîche	Normale	Fort	Non
Couvert	Fraîche	Normale	Fort	Oui
Ensoleillé	Tiède	Elevée	Faible	Non
Ensoleillé	Fraîche	Normale	Faible	Oui
Pluie	Tiède	Normale	Faible	Oui
Ensoleillé	Tiède	Normale	Fort	Oui
Couvert	Tiède	Elevée	Fort	Oui
Couvert	Chaude	Normale	Faible	Oui
Pluie	Tiède	Elevée	Fort	Non

**Apprentissage Supervisé.** Le but de l'apprentissage supervisé est de construire un modèle de prédiction, appelé aussi classifieur, qui nous permettra d'identifier un attribut à prédire  $Y$ , appelé variable endogène, classe, variable à expliquer, variable à prédire, à partir d'un certain nombre d'attributs explicatifs appelés variables exogènes, variables explicatives, variables prédictives, notées  $X$  [11].

- Le modèle de prédiction ou fonction de classement  $\varphi$  est construit sur un sous ensemble de la population, appelé échantillon d'apprentissage noté  $\Omega_a$ .
- Un individu appartenant à l'échantillon est noté  $\omega$ .

L'attribut à prédire  $Y$  associé à chaque individu de  $\Omega_a$  une classe appartenant à  $C$  {ensemble des classes  $C = \{c_1, \dots, c_m\}$ },  $Y : \Omega_a \rightarrow C$

$$\omega \rightarrow Y(\omega)$$

- Chaque variable exogène (explicative)  $X_j$  est définie par :  $X_j : \Omega_a \rightarrow E_j$

$$\omega \rightarrow X_j(\omega)$$

$E_j$  : ensemble des modalités (valeurs) de  $X_j$ .

Parmi les méthodes de l'apprentissage supervisé, on s'intéresse plus particulièrement aux méthodes à base d'arbres de décision.

**Méthodes à Base d'Arbres de Décision.** Les modèles de classification générés par ces méthodes sont des arbres, un arbre est constitué de plusieurs niveaux appelés partitions  $S_i$ , une partition est composée de nœuds ou sommets  $s_j$  qui sont reliés par des arcs. La génération de l'arbre consiste à la génération de partitions, une partition représente un ensemble de sommets générés à un instant  $t$ , le passage de  $S_i$  à  $S_{i+1}$  se fait en optimisant un critère de variation d'incertitude  $\tau(S_{i+1}) = \tau(S_i) - \tau(S_{i+1})$ . Parmi les méthodes, on retrouve ID3 [21], C4.5 [22,23], CART [7], SIPINA [1,2,27], etc. SIPINA est une méthode qui génère un graphe et non un arbre du fait de la présence de cycles.

1. Extraction des règles [24,25]

A partir d'un arbre de décision, il est possible d'extraire des règles de la forme :

Si <Conditions> Alors <Classe= valeur>.



A partir de l'arbre de la figure ci-dessus, on peut extraire 05 règles correspondant aux feuilles :

- Règle 1 : Si Ciel='Ensoleillé' ET Humidité='Elevée' Alors Jouer\_Tennis=Non.
- Règle 2 : Si Ciel='Ensoleillé' ET Humidité='Normale' Alors Jouer\_Tennis=Oui.
- Règle 3 : Si Ciel='Couvert' Alors Jouer\_Tennis=Oui.
- Règle 4 : Si Ciel='Pluie' ET Vent='Fort' Alors Jouer\_Tennis=Non.
- Règle 5 : Si Ciel='Pluie' ET Vent='Faible' Alors Jouer\_Tennis=Oui.

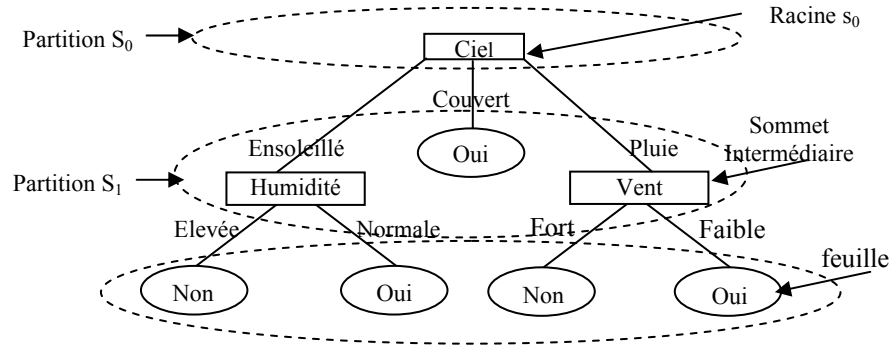


Fig. 2. Arbre de décision de l'échantillon du tableau 1 généré par ID3.

**Transformation de l'Arbre de Décision dans le Formalisme d'Automates d'Arbres.** L'automate d'arbres calculé  $A$  est défini par le quadruplet  $(Q, V, \Delta, Q_f)$  tel que :  $V$  est un élément de  $\Sigma_S$  et  $\Delta$  est un ensemble qui contient les règles de transition  $\delta$ , son calcul est réalisé au fur et à mesure de la génération de l'arbre.

L'opération de base de construction d'un arbre de décision est l'éclatement, on a deux types d'éclatements à considérer et à traduire dans le formalisme d'automate :

- ✓ Eclatement du nœud racine

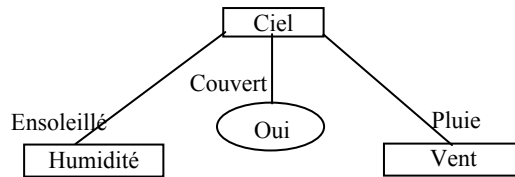


Fig. 3. Arbre de décision du tableau 1 généré à un instant  $t_1$ .

L'automate d'arbre  $A = (Q, V, \Delta, Q_f)$  correspondant à l'arbre généré à un instant  $t_1$  est composé de :  $Q_f = \{q_0, q_1\}$ ,  $Q_f$  contient initialement les états  $q_0$  correspondant à la modalité Non de la classe 'Jouer\_Tennis' et  $q_1$  correspondant à la modalité Oui,  $Q = Q_f \cup \{q_{abs}, q_2, \dots, q_{10}\}$ ,  $q_{abs}$  correspond à l'état absent,  $V = V \cup \{Ciel, (., Ensoleillé, Couvert, Pluie)\}$ ,  $\Delta = \{Absent \rightarrow q_{abs}, Ensoleillé \rightarrow q_2, Couvert \rightarrow q_3, Pluie \rightarrow q_4, Ciel(q_2, q_5) \rightarrow q_8, Ciel(q_3, q_6) \rightarrow q_9, q_{abs} \rightarrow q_6, Ciel(q_4, q_7) \rightarrow q_{10}, q_9 \rightarrow q_1\}$ ,

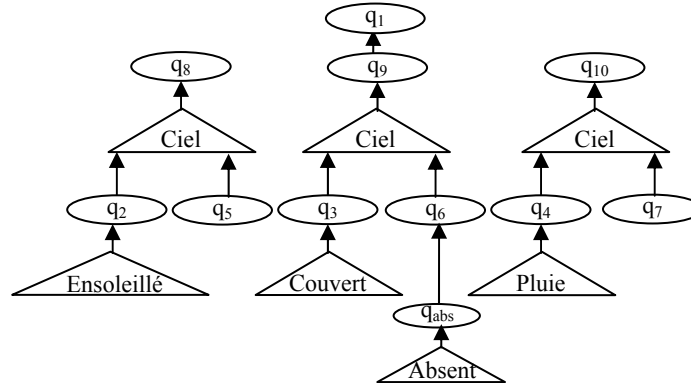


Fig. 4. Automate d'arbre correspondant à l'arbre de la figure 3.



L'automate d'arbre  $A = (Q, V, \Delta, Q_f)$  correspondant à l'arbre généré est composé de :  $Q = \{q_{abs}, q_0, \dots, q_{22}\}$ ,  $V = V \cup \{\text{Vent } (,), \text{Fort, Faible}\}$ ,  $\Delta = \Delta \cup \{\text{Fort} \rightarrow q_{17}, \text{Vent}(q_{17}, q_{18}) \rightarrow q_7, \text{Faible} \rightarrow q_{19}, \text{Vent}(q_{19}, q_{20}) \rightarrow q_{21}, \text{Ciel}(q_{21}, q_4) \rightarrow q_{22}, q_{abs} \rightarrow q_{12}, q_8 \rightarrow q_0, q_{abs} \rightarrow q_{14}, q_{16} \rightarrow q_1, q_{abs} \rightarrow q_{18}, q_{10} \rightarrow q_0, q_{abs} \rightarrow q_{20}, q_{22} \rightarrow q_1\}$ .

✓ Extraction des règles

Les règles extraites de l'arbre de décision de la figure 2, peuvent aussi être extraites de l'automate de la figure 7. Ceci est réalisé à partir de l'ensemble  $\Delta$  en considérant les règles- $\varepsilon$  de la forme  $q_i \rightarrow q_0$  ou  $q_i \rightarrow q_1$  ;  $q_i \in Q - \{Q_f\}$  et  $q_0, q_1 \in Q_f$ .

Chaque règle- $\varepsilon$  correspond à une règle de l'arbre. En effectuant des substitutions dans les règles- $\varepsilon$  on obtient les cinq règles suivantes :

- 1)-  $q_{16} \rightarrow q_1 \Rightarrow \text{Ciel}(q_{15}, q_2) \rightarrow q_1 \Rightarrow \text{Ciel}(\text{Humidité}(q_{13}, q_{14}), q_2) \rightarrow q_1 \Rightarrow$  Si Ciel='Ensoleillé' et Humidité='Normale' Alors Jouer-Tennis='Oui',
- 2)-  $q_9 \rightarrow q_1 \Rightarrow \text{Ciel}(q_3, q_6) \rightarrow q_1 \Rightarrow$  Si Ciel='Couvert' Alors Jouer-Tennis='Oui',
- 3)-  $q_{10} \rightarrow q_0 \Rightarrow \text{Ciel}(q_4, q_7) \rightarrow q_0 \Rightarrow$  Si Ciel='pluie' et Vent='Fort' Alors Jouer-Tennis='Non',
- 4)-  $q_{22} \rightarrow q_1 \Rightarrow \text{Ciel}(q_{21}, q_4) \rightarrow q_1 \Rightarrow \text{Ciel}(\text{Vent}(q_{19}, q_{20}), q_4) \rightarrow q_1 \Rightarrow$  Si Ciel='pluie' et Vent='Faible' Alors Jouer-Tennis='Oui',
- 5)-  $q_8 \rightarrow q_0 \Rightarrow \text{Ciel}(q_2, q_5) \rightarrow q_0 \Rightarrow$  Si Ciel='Ensoleillé' et Humidité='Elevée' Alors Jouer-Tennis='Non'.

## 4 Conclusion

L'objectif de cet article est de présenter la problématique de notre travail et qui concerne la conception et l'expérimentation d'une nouvelle technique de génération et d'optimisation d'arbres de décision par une méthode formelle de modélisation : les automates d'arbres. Nous nous sommes fixés comme objectifs l'expérimentation des algorithmes de minimisation d'automates d'arbres pour simplifier les modèles de classification générés par des techniques à base d'arbres de décisions : ID3 [21], C4.5 [22,23], CART [7], SIPINA [1,2,27]. Notre perspective concerne l'intégration de la simplification des bases de règles dans une interface d'acquisition automatique des connaissances à partir des données. Cette simplification a pour objective la détection des incohérences dans les bases de règles et l'élimination des variables exogènes non pertinentes.

## Références

1. Atmani, B., Beldjilali, B.: Neuro-IG : A Hybrid System for Selection and Elimination of Predictor Variables and non Relevant Individuals, Informatica, Journal International, Vol. 18, N°2 (2007) 163-186.
2. Atmani, B., Beldjilali, B.: Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, Vol.26, N°2 (2007) 171-197.
3. Bernard, M., De La Higuera, C.: GIFT "Grammatical Inference For Terms". Late Breaking paper at the 9<sup>th</sup> International Conference on Inductive Logic Programming (1999).
4. Bernard, M., De La Higuera, C.: Apprentissage de programmes logiques par inférence grammaticale. Revue d'Intelligence Artificielle, 14(3/4), (2001) 375-396
5. Bouchou, B., Halfeld Ferrari Alves, M., Musicante, M.: Tree automata to verify XML key constraints, 6th International Workshop on the Web and Data Bases WebDB (2003) 37-43
6. Bouchou, B., Halfeld Ferrari Alves, M.: Updates and Incremental Validation of XML Documents, 9th Int. Conference on Database Programming Languages, in LNCS 2921/2003 DBPL (2003) 216-232
7. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification And Regression Trees, Chapman et Hall (1984)
8. Carne, J., Gilleron, R., Lemay, A., Niehren, J.: Interactive learning of node selecting tree transducers. Machine Learning (2006)
9. Carrasco, R., Oncina, J., Calera-Rubio, J.: Stochastic Inference of Regular Tree Languages. Machine Learning, 44(1/2), (2001) 185-197
10. Comon, H., Dauchet, M., Gilleron, R., Jacquemard, F., Leguiez, D., Tison, S., Tommasi, M. : Tree Automata Techniques and applications (2005)
11. Denis, F., Gilleron, R.: Apprentissage à partir d'exemples, Université Charles de Gaulles, Lille (1997)
12. Genet, T., Klay, F.: Rewriting for Cryptographic Protocol Verification in Proceedings 17<sup>th</sup> International Conference on Automated Deduction, série Lecture Notes in Artificial Intelligence, volume 1831, Springer-Vergla (2000)
13. Genet, T., Viet Tiem Tong, V.: Reachability Analysis of Term Rewriting Systems with Timbuk. in « Proceedings of the 8th International Conference on Logic for Programming, Artificial Intelligence and Reasoning », série Lecture Notes in Artificial Intelligence volume 2250. Springer-Verlag (2001) 691-702

14. Genet, T., Tang-Talpin, Y. M., Viet Triem Tong, V.: Verification of copy-protection cryptographic protocol using approximations of term rewriting systems. In Proc. Of WITS'03, Workshop on Issues in the Theory of Security (2003)
15. Goubault-Larrecq, J.: A method for automatic cryptographic protocol verification (extended abstract). In Proceedings of the International Workshop on Formal Methods in Parallel Programming, Techniques and Applications, volume 1800. Springer Verlag Lecture Notes in Computer Science (2000) 977–984
16. Habrard, A., Bernard, M., Jacquenet, F.: Apprentissage d'automates d'arbres stochastiques généralisés à partir de bases de données relationnelles. In Conférence d'Apprentissage, Orléans, France, Presses Universitaires de Grenoble (2002) 180–191
17. Habrard, A.: Modèles et techniques en inférence grammaticale probabiliste : de la gestion du bruit à l'extraction de connaissances Octobre (2004)
18. Monniaux, D.: Abstracting Cryptographic Protocols with Tree Automata, Static Analysis Symposium (1999)
19. Oncina, J., Garcia, P.: Inferring regular languages in polynomial update time. In Pattern Recognition and Image Analysis (1992) 49-61
20. Preux, P.: Fouille de données, Notes de cours. Université de Lille 3, Juillet (2006)
21. Quinlan, J.R.: Induction of Decision Trees, Machine Learning 1 (1986) 81-106
22. Quinlan, J.R.: Probabilistic decision trees. In Machine Learning: An Artificial Intelligence Approach, Volume 3. Morgan Kaufmann (1990)
23. Quinlan, J.R.: C4.5 Programs for Machine Learning, Morgan Kaufmann (1993)
24. Rabaseda, S., Rakotomalala, R., Sebban, M.: Génération automatique de connaissances par induction. Actes des 3<sup>èmes</sup> rencontres de la société francophone de classification (1995) 45-46
25. Rabaseda, S., Rakotomalala, R., Zighed, D. A.: Rules extracted automatically by induction. Proceeding of the 6<sup>th</sup> conference on information processing and management of uncertainty (1996) 551-556
26. Tommasi, M.: Structures arborescentes et apprentissage automatique. Habilitation à diriger des recherches, Université de Lille 3, France Novembre (2006)
27. Zighed, D. A., Auray, J. P., Duru, G.: SIPINA : Méthode et Logiciel. Lacassagne (1992)
28. Dal Zilio, S., Lugiez, D.: Fondements de l'interrogation de données semi structurées. Laboratoire d'Informatique Fondamentale de Marseille, CNRS, UMR 6166, ATIP Jeunes chercheurs (2002)
29. Dal-Zilio, S., Lugiez, D.: XML Schema, Tree Logic and Sheaves Automata. Rapport technique, INRIA (2002)
30. Dal Zilio, S., Lugiez, D.: XML Schema, Tree Logic and Sheaves Automata. In Proc. of RTA - Rewriting Techniques and Applications, LNCS 2706 (2003)
31. Dal Zilio, S., Acciai, L.: Pattern matching et documents XML “Un nouvel opérateur de filtrage pour les documents XML (2004)
32. Dal Zilio, S., Lugiez, D., Meyssonier, C.: Logic you Can Count On. In Proc. of the 31<sup>st</sup> ACM SIGPLAN-SIGACT symposium on Principle of Programming Languages (2004) 135-146.

# Vérification et Reconnaissance des signatures manuscrite hors-ligne en utilisant les SVMs Multi-classes et les Réseaux De Neurons A Impulsion

BELHIA .S<sup>1</sup>, BENYETTOU.M<sup>2</sup>, LEHIRECHE.A<sup>1</sup>, ADJOUDJ.R<sup>1</sup>

<sup>1</sup> Laboratoire EEDIS,, Université Djillali Liabès De Sidi bel-Abbès.

<sup>2</sup>Laboratoire LAMOSI, USTO, Université des Sciences et Technologies d'Oran.

Belhia\_souad@hotmail.com, elhir@univ-sba.dz, adjreda@yahoo.fr.

**Abstract.** Ce document présente un système de vérification et de reconnaissance de la signature manuscrite hors-ligne, qui utilise l'approche de la Machine à vecteurs de support (SVM<sup>3</sup>) multi-classes dans le but de vérifier et de reconnaître les signatures, en se basant sur la stratégie « *one against one* » et « *one against all* ». En comparant en suite les performances de cette approche avec celles de l'approche neuronal à « *impulsion* » ou à « *Spike* », (les SNNs<sup>4</sup>), cette dernière utilise trois types d'apprentissage: Back-propagation, Analyse discriminante de Fisher et les règles *Pdelta*<sup>5</sup>.

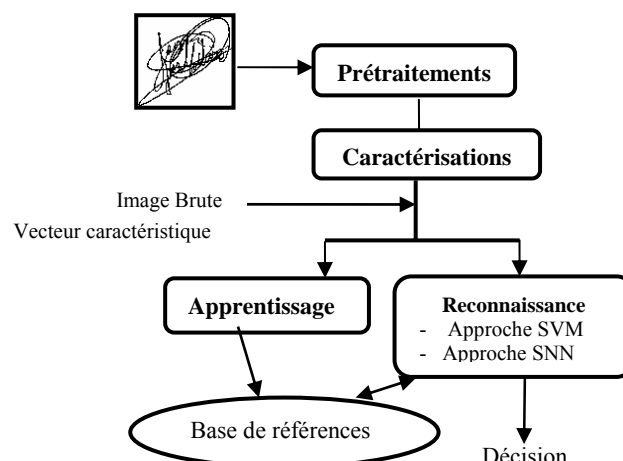
**Mots clés :** Signature manuscrite, Apprentissage automatique, Biométrie, la Machine à vecteurs de support, Réseau de neurone à impulsion.

## 1. Introduction :

La reconnaissance et la vérification de la signature est parmi les axes les plus importants dans le domaine de la biométrie. En effet, la signature a toujours été le moyen le plus accepté socialement et légalement pour l'identification et l'authentification de tout document officiel. Les signatures sont caractérisées par une grande variabilité intra-classe. De plus, la signature authentique en général possède une diversité infinie de styles. Pour cela, Différentes approches ont été proposées pour l'authentification hors ligne de la signature,

## 2. Principes de fonctionnement :

Classiquement, la conception d'un système d'authentification nécessite d'apporter des solutions à quatre problèmes :



**Fig (1)** Architecture d'un système de reconnaissance de signature hors-ligne

<sup>3</sup> Support Vector Machine.

<sup>4</sup> Spike Neural Network.

<sup>5</sup> Parallel Delta Rule,

**2.1. Acquisition des données:** La signature est digitalisée par un scanner. C'est l'entrée de ce système. La résolution normale est 300 dpi. Pourtant, quand la taille de la signature est petite, il faut augmenter la résolution. Les bases de données utilisées la première base de données, la feuille est scannée sous la résolution de 75 dpi, et la deuxième sous la résolution normale de 300 dpi.

**2.2. Prétraitement :** Cette étape prépare l'image d'entrée pour l'étape de reconnaissance, elle essaie d'effacer les informations indésirables comme le bruit. Les opérations typiques sont, le seuillage, l'élimination du bruit et la normalisation, pour simplifier la comparaison entre les signatures, la taille des images est de (128x128)- Fig 1.

**2.3. Caractérisations:** nous avons utilisé les filtres de Gabor [17]. Après avoir appliqué la transformation de Gabor sur une image, une région de texture est caractérisée par la moyenne et la variance des coefficients de transformation, un vecteur de caractéristiques est construit en utilisant ces caractéristiques comme composants.-Fig 1.

**2.4. Apprentissage et Reconnaissance :** pour cela nous avons implémenté les algorithmes suivants : Les SVM multi-classes selon les stratégies "one against one" et "one against all" [Fig 2-3] et Le Réseau de neurone à impulsion Fig 4.

### 3. Apprentissage :

Deux paradigmes ont été utilisés le paradigme statistique et le paradigme connexionniste.

#### 3.1 Description de l'Approche basée sur la Théorie des « Support Vector Machines »

**3.1.1 Principe:** Depuis quelques années, des nouvelles méthodes d'apprentissage se développent sur la base de la Théorie de l'Apprentissage Statistique de Vapnik et Chervonenkis [1]. L'une de ces méthodes, appelée Machine à Vecteur de Support ou SVM (*Support Vector Machine*), permet de réaliser des estimations en classification [2] ou en régression [3].

**Le Cas de deux classes :** Cette approche est basée sur un algorithme d'apprentissage ayant comme fonction de décision un hyperplan optimal dans un espace de grande dimension que nous allons appeler  $F$ . A partir des données d'apprentissage  $\{x_i, y_i\}$  Tq  $i=1, \dots, n$ ; où  $x_i \in R_n$  et  $y_i \in \{-1, 1\}$ , l'algorithme consiste à représenter les points  $x_i$  dans cet espace  $F$  en utilisant une transformation non-linéaire, construite à partir d'une famille de fonctions  $\{\Phi_k\}_{k \in N}$ .  $F$  peut être vu comme l'espace vectoriel généré par cette famille  $\{\Phi_k\}$ . Sous certaines conditions très générales, il se trouve que la série  $\sum_{k \in N} \Phi_k(x) \Phi_k(y)$  converge vers une fonction de  $F$  appelée le noyau, que l'on note :

$$k(x, y) = \sum_{k \in N} \Phi_k(x) \Phi_k(y) \quad (1)$$

Dans ce cas, la fonction de décision est donnée par le signe de la fonction de discrimination suivante qui ne dépend plus que du noyau  $K$  :

$$f(x) = \sum_{i=1}^l y_i \alpha_i K(x, x_i) + b \quad (2)$$

Où les  $\alpha_i$  et  $b$ , sont des coefficients à déterminer, en maximisant la distance, appelée *marge*, entre la frontière de décision  $f(x) = 0$  et le nuage de point dans  $F$ .

$$\begin{cases} \min_{f \in F} \|f\|_F^2 + C \sum_{i=1}^l |\xi_i|, \\ \text{avec } y_i f(x_i) \geq 1 - \xi_i, i = 1, l \\ \text{et } \xi_i \geq 0, i = 1, l \end{cases}$$

Avec  $C$  un paramètre qui permet de régler le taux d'erreur admissible dans la solution (si  $C = 0$ , on peut faire autant d'erreurs que possible, si  $C = 1$ , on n'admet aucune erreur). En pratique, ce problème d'optimisation se résout de la manière suivante :

$$\begin{cases} \max_{\alpha_i} \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{avec } 0 \leq \alpha_i \leq C, i = 1, l \\ \text{et } \sum_{i=1}^l \alpha_i y_i = 0. \end{cases}$$

On peut se demander comment déterminer  $K(x, y)$  à partir des fonctions  $k$ , mais on pose en fait le problème dans l'autre sens : pour un noyau donné  $K(x, y)$ , existe-t-il une famille des  $k$  vérifiant la fonction (1) Pour vérifier cela, il est nécessaire et suffisant que le noyau  $K(x, y)$  soit symétrique et semi défini positif, c'est-à-dire que :

$$\{\forall (\alpha_i)_{i \in N} \in \mathbb{R}, \forall (z_i)_{i \in N} \in \mathbb{R}^n, \sum_i \sum_j \alpha_i \alpha_j k(z_i, z_j) \geq 0\}$$

**Le Cas multi classes :** L'utilisation des SVMs dans le cas multi classe peut se faire de deux façons différentes: par stratégie de type « un contre les autres » ou « un contre un » [5].

**3.1.1.1 (one against one) :** Pour cette stratégie, on apprend  $k$  ( $k - 1$ )/2 classificateurs où chaque classificateur apprend en fait à discriminer les points de deux classes différentes. Dans ce cas, l'affectation d'une classe à un nouveau point  $x_t$  consiste à évaluer la classe associée à ce point par les différentes fonctions de décision et de décider ensuite par une règle de vote.

**3.1.1.2 (one against all) :**

Pour cela  $k$  frontières de décision  $f_k(x)$  séparant une classe des autres sont générées.  $x$  est affecté à la classe  $v$  si :

$$v = \arg \max_k f_k(x)$$

L'architecture du système en stratégie Un-Contre-Tous Fig(3)<sup>6</sup>. Notons, que l'étage étiqueté «fusion» désigne le schéma de vote utilisé et aussi toute sorte d'expert capable de fusionner les sorties pour décider de la classe d'appartenance.

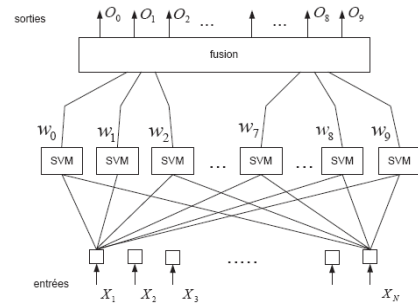


Fig (2) Architecture du système en stratégie one against one

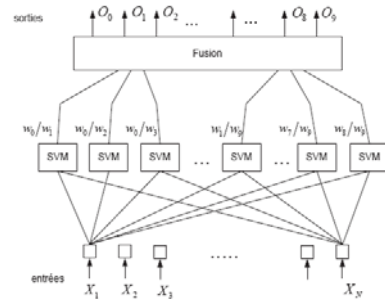


Fig (3) Architecture du système en stratégie one against all

## 3.2 Description de l'Approche basée sur les SNNs

**3.2.1 Principe :** Le *Spiking Neural Networks* (SNN) est un modèle abstrait artificiel qui simule le système nerveux humain, les neurones biologiques communiquent entre eux en utilisant des impulsions de courte durée, appelées potentiels d'action ou *Spikes*, ces derniers se sont des signaux caractérisés par une amplitude et des propriétés temporaires. En plus, les SNNs utilisent les *Spikes* au lieu de la fonction de transfert *sigmoïde* que les ANNs<sup>7</sup> à perceptron multi-couche utilisent.

**3.2.2 Le modèle "Integrate-and-Fire" :** Le modèle "Integrate-and-Fire" [6], est le modèle le plus simples : son activité interne est mesurée simplement par son potentiel de membrane  $V$ , et à l'image du neurone biologique, comprend une partie intégratrice (l'arbre dendritique), qui va correspondre à une somme linéaire des entrées pondérée par les poids synaptiques et une partie non-linéaire qui simule la génération d'un PA<sup>8</sup> quand  $V$  dépasse un seuil : on dit alors que le neurone "tire". Celui-ci est alors inhibé (pendant la période réfractaire) en fixant durant cette période son potentiel à une valeur dépolarisée. Par souci de réalisme, on peut aussi introduire un courant de fuite pour obtenir le modèle LIF ( *Leaky Integrate-and-Fire*). Par analogie avec un circuit électrique du premier ordre, le LIF est défini par une équation différentielle du 1<sup>er</sup> ordre de constante de temps  $\tau_m = \frac{C_m}{g_m}$ ,

<sup>6</sup> ( $O_i$  output du SVM.  $W_i$  la probabilité que  $\S$  appartienne à classes  $i$ .  $X_{1m}$  les inputs.)

<sup>7</sup> Artificial Neural Network.

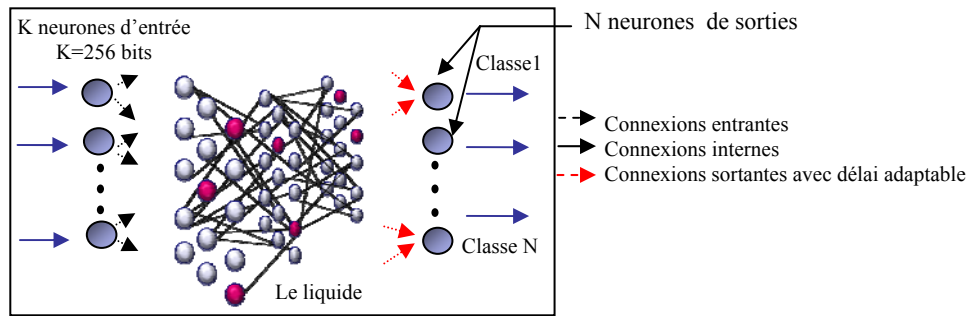
<sup>8</sup> Potentiel d'Action.

où  $g_m$  est la conductance,  $c_m$  est la capacité membranaire, caractérisant le temps de fuite. Le potentiel de membrane  $V(t)$  dévie alors du potentiel de repos  $V_{repos}$  en intégrant le courant d'entrée  $I(t)$  jusqu'à atteindre le seuil. La sortie est alors la suite des PAs.

$$\left\{ \begin{array}{l} C_m \frac{dV_i(t)}{dt} = -g_m (V_i(t) - V_{repos}) + I_i(t) \\ V_i(t) > V_{seuil} \text{ Alors } \left\{ \begin{array}{l} t_i^{(f)} := t \\ \text{Et } V_i(t_i^k < t < t_i^k + t_{ref}) = V_{RAZ} \end{array} \right. \end{array} \right.$$

Les potentiels  $V_{seuil}$  et  $V_{RAZ}$  correspondent respectivement aux potentiels de seuil et de remise à zéro.

**3.2.3 Le modèle "Liquide State Machine" :** L'implémentation est inspirée d'un simulateur appelé LSM (*Liquid State Machine*) proposé par Mass et al [8] ; Qui peut modéliser un grand réseau hétérogène qui consiste en plusieurs modèles de neurones et synapses différents. La dimension de ce réseau est jusqu'à 10 000 neurones avec quelques millions de synapses.



Fig(4) architecture de réseau proposé.

Le modèle est composé d'un réseau principal de  $M$  neurones internes, d'un ensemble de  $K=256$  neurones d'entrée, et d'un ensemble de sorties qui comporte  $N$  neurones de sortie, chaque neurone représente une classe.

**L'ensemble d'entrée :** Permet de présenter une entrée au réseau, elle est utilisée pour transformer des données analogiques en trains d'impulsions.

**Le réseau principale « filtre liquide » :** réalise une mémorisation des entrées ainsi qu'une projection dans un espace de dimension élevée. Il n'a pas d'organisation préconçue et sa connectivité interne est partielle ; nous appelons ce réseau principal « le filtre liquide » reflétant ainsi l'absence d'une structure prédéfinie, la seule donnée imposée est la proportion entre neurones excitateurs et Inhibiteurs. Pour respecter les observations biologiques, nous imposons 80% de neurones excitateurs et 20% de neurones inhibiteurs. Chaque connexion synaptique d'un neurone  $N_i$  vers un neurone  $N_j$  est caractérisée par un poids  $w_{ij}$  et un délai axonal  $d_{ij}$ . Le filtre liquide est un réseau fortement récurrent de neurone impulsionnels, de type *Integrate-and-Fire*, il joue le rôle d'un réservoir de dynamique [10].

**L'ensemble de sortie :** est une carte de lecture, qui permet d'extraire de l'état du réseau à un instant donné une information déterminée par apprentissage sur la carte de lecture.

### 3.2.4 Les algorithmes d'apprentissage

**3.2.4.1 La retro-propagation :** Est une méthode qui permet de calculer le gradient de l'erreur pour chaque neurone du réseau, de la dernière couche vers la première.

**3.2.4.2 Apprentissage par l'Analyse Discriminante de Fisher :** La technique de Fisher permet une réduction dans le nombre de dimensions,  $d$  avec une faible augmentation dans la probabilité d'erreur

**Algorithme :**  $C$  classes,  $N_c$  exemples dans la classe  $c$ , nombre total :

Pour chaque classe  $c$  on estime la moyenne :  $\hat{m}_c, \hat{\Sigma}_c = \sum_{i=1}^{N_c} x_i x_i^T$   
 $x \in \mathbb{R}^d$ , on recherche une nouvelle variable :  $v$   
 Résumant au mieux les  $d$  variables observées

Inertie INTER classes :  $S_B = \sum_{c=1}^C \frac{N_c}{n} (\hat{m}_c - m)(\hat{m}_c - m)^T$



$$\text{Inertie INTRA classes : } S_W = \sum_{c=1}^C \frac{n_c \hat{\Sigma}_c}{n}$$

$$\min_{v \in \mathbb{R}^d} \frac{v^T S_B v}{v^T S_W v} \quad \text{Solution : } v=0 \text{ ou } \bullet$$

$$\text{On cherche une direction : } \min_{v \in \mathbb{R}^d} \frac{v^T S_B v}{v^T S_W v} \quad \text{avec } v^T S_W v = 1$$

$$\begin{aligned} & \min_{v \in \mathbb{R}^d} \frac{v^T S_B v}{v^T S_W v} \quad \text{avec } v^T S_W v = 1 \\ & \Leftrightarrow \min_{v \in \mathbb{R}^d} v^T S_B v \quad \text{avec } v^T S_W v = 1 \\ & \Leftrightarrow \min_{v \in \mathbb{R}^d} v^T S_B v - \lambda (v^T S_W v - 1) \\ & \Leftrightarrow \frac{\partial (v^T S_B v - \lambda (v^T S_W v - 1))}{\partial v} = 0 \\ & \Leftrightarrow S_B v - \lambda S_W v = 0 \\ & \Leftrightarrow S_W^{-1} S_B v = \lambda v \end{aligned}$$

On cherche les vecteurs propres de la matrice  $S_W^{-1} S_B$

**3.2.4.3 Parallel delta rule :** Récemment Awer [10] a introduit le “*p-delta rule*” pour les perceptrons parallèles, les réseaux de neurones à une seule couche de perceptrons. Le “*p-delta rule*” est une règle utilisée pour la mise à jour des poids des neurones dans une seule couche de perceptrons.

**Algorithme:** Pour le neurone  $j$  avec la fonction d'activation  $g(x)$ .

- La règle delta pour le  $j^{\text{ème}}$  poids  $w_{ji}$  est donnée par :  $\bullet w_{ji} = \bullet (t_j - x_j) g'(h_j) x_i$ , où  $\bullet$  est une petite constante,  $g(x)$  est la fonction d'activation du neurone,  $t_j$  est le label de la sortie,  $x_j$  est la sortie réelle et  $x_i$  est la  $i^{\text{ème}}$  entrée.

- La règle p-delta est présentée pour un perceptron avec une fonction d'activation linéaire comme :  $\bullet w_{ji} = \bullet (t_j - x_j) x_i$

- La règle delta est dérivée, en essayant de minimiser l'erreur dans la sortie du perceptron à travers la descente du gradient. Les erreurs pour un perceptron avec les  $j$  sorties peuvent être mesurées par :  $E = \bullet 1 / 2 (t_j - x_j)^2$

- La dérivation par rapport aux poids s'écrit comme :

$$\text{- La dernière équation s'écrit : } \frac{\partial E}{\partial w_{ij}} = -(t_j - x_j) g'(h_j) x_i$$

Choisir une proportionnalité  $S$  constante et éliminer le signe moins nous permettra de déplacer le poids dans la direction négative du gradient et minimiser l'erreur, nous arrivons à l'équation de cible: [12]

$$\bullet w_{ji} = \bullet (t_j - x_j) g'(h_j) x_i$$

**4. La Base de données :** Les différentes expérimentations ont été effectuées sur deux bases de données.



Fig(5) Quelques échantillons de la deuxième base de données d'extension (.png) [11].



Fig(6) : Quelques échantillons de la première base de données d'extension (.bmp) [11]

- **La première base de données** [11] : est constituée de 8640 signatures authentiques de format « .png » correspondant à 160 signataires, chacun ayant fourni 48 échantillons. 3840 signatures (24 échantillons correspondant à chaque signataire), représente la base d'apprentissage ; et 4800 signatures, (30 échantillons de chaque type) pour la base de test. Chaque personne représente une classe-Fig 5.

- **la deuxième base de données** [11] : contiennent 7680 signatures originales de format « .bmp ». 3840 signatures (24 échantillons correspondant à chaque signataire), représente la base d'apprentissage et 3840 signatures authentiques, (24 échantillons de chaque type) pour la base de test. Fig 6.

**4.1 Résultats expérimentaux de SVM multi-classes :** Deux méthodes des combinaisons binaires “*One against one*” et “*One against all*”, décrits précédemment, ont été comparées expérimentalement.

**4.2 Discussions et Comparaison entre les approches multi-classes « un contre un » et « un contre tous » :** On peut constater que les résultats obtenus à partir des méthodes SVMs-Multi-classes, « one against one » et « one against all », sont identiques par rapport à la première base de données, mais le temps d'apprentissage diffère, la méthode « one against all » est plus rapide à celle de la méthode « one against one ».

#### 4.3 Résultats expérimentaux de réseau de neurone à impulsion

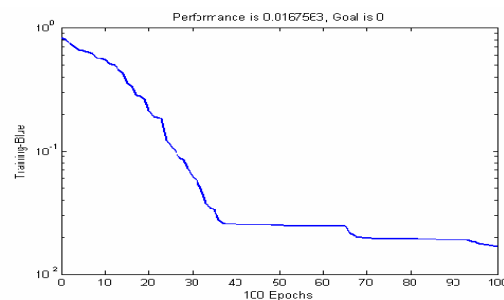
Les tableaux ci dessous résument les résultats expérimentaux en utilisant les trois algorithmes d'apprentissage :

**Tableau. 1** Résultats expérimentaux obtenus par les deux méthodes « un contre un » et « un contre Tous »

	One against one			One against all		
	Temps d'app	Temps de test	TR <sup>9</sup>	Temps d'app	Temps de test	TR
BDD 1 (.png)	30.37 mnt	2.27 mnt	100%	2.54 mnt	13.30 sec	100%
BDD 2 (.bmp)	16.43 mnt	2.40 mnt	95.6%	3.41 mnt	9.82 sec	51.83%

**Tableau .2** Représente les résultats obtenus lors de l'utilisation de **retro-propagation**

Réseau de neurone à impulsion				
Apprentissage par rétro-propagation				
	Temps d'app	Temps de test	TR	MSE
BDD1	29.589 mnt	0.07sec	98.75%	0.0125
BDD 2	29.60 mnt	0.50sec	92.43%	0.0756



**Fig.7.** La convergence du SNN après 100 itérations (epochs)

**Tableau.3** les résultats obtenus lors de l'utilisation de l'approche d'Analyse Discriminante de Fisher

Réseau de neurone à impulsion				
Apprentissage par analyse discriminante de fichier				
	Temps d'app	Temps de test	TR	Erreur
BDD 1	6.21 sec	0.26sec	87.66%	0.123
BDD 2	7.171sec	0.03sec	85.72%	0.142

**Tableau.4** Représente les résultats obtenus lors de l'utilisation de l'approche p-delta

Réseau de Neurone à Impulsion				
Apprentissage par P-delta				
	Temps d'app	Temps de test	TR	Erreur
BDD 1	0.04 sec	0,02 sec	81,50%	0.185
BDD2	0,03 sec	0,03 sec	96,89%	0.031

**4.4 Comparaison entre les trois algorithmes d'apprentissages utilisés dans le Réseau de Neurones à Spike :** Nous pouvons voir dans le tableau (2), que l'algorithme Retro-propagation donne de meilleur résultat, 98.75% pour la première base de données. La deuxième BDD a donné 92.43% avec un temps d'apprentissage très lent. Tandis que l'approche Analyse Discriminante de Fisher est très rapide avec un taux de reconnaissance moins performant. 87.66% entre 85.72. Pour le troisième algorithme d'apprentissage « le p-delta » a donné un taux de reconnaissance de 84.38% pour le premier jeu de données et le meilleur taux de reconnaissance pour le deuxième jeu de données.

<sup>9</sup> TR : taux de reconnaissance

## 5. Conclusion

Les approches proposées ont montré que le problème de la reconnaissance des signatures manuscrites peut être résolu efficacement en utilisant une approche à base des SVMs multi-classes et les SNNs, elle est basée sur les modèles des signatures, qui sont considérés comme des exemples d'apprentissage pour les SVMs et les SNNs. Le réseau de neurones impulsionsnels ouvre la voie à de nouveaux développements théoriques dans la modélisation des modalités sensorielles. Le simulateur LSM nous a permis d'évaluer les performances qu'on pouvait attendre d'une telle approche, combinant une assez large flexibilité de modélisation, Pour le test nous avons eu aussi des résultats favorables.

## Référence

1. Vapnik Support vector learning machines. Tutorial at NIPS\*97, Denver (CO), Decembre 1997
2. C. Burges. A tutorial on support vector machine for pattern recognition, 1998.
3. A., Smola. *Learning with Kernels*. PhD thesis, GMD First, Berlin, Germany, 1998.
4. CANU S., MARY X. & RAKOTOMAMONJY A. (2003). *Advances in Learning Theory : Methods, Models and Applications NATO Science Series III : Computer and Systems Sciences*, chapter Functional learning through kernel. IOS Press, Amsterdam.
5. Weston J. & Watkins C. (1999). Support vector machines for multiclass pattern recognition. In *Proceedings of the Seventh European Symposium On Artificial Neural Networks*.
6. Lapique L (1907) recherche quantitative sur l'excitation électrique des nerfs traités comme une polarisation J. physiol. Pathol. Gen. 9, pp. 620-635.
7. Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560.
8. Maass, W., Natschläger, T., and Markram, H. (2002b). A fresh look at real time computation generic recurrent neural circuits, submitted for publication.
9. Olivier Rochel, Une approche événementielle pour la modélisation et la simulation de réseaux de neurones impulsionsnels 2004.
10. H. Jaeger The "echo state approach to analysing and training recurrent neural networks", technical report GMD report 148 german National research center for information technology, 2001.
11. Miguel A. Ferrer, Jesús B. Alonso, Carlos M. Travieso, "Offline geometric parameters for automatic signature verification using fixed-point arithmetic", in *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 6, pp 993-997, June 2005.
12. Zebin Jiang, "Liquid State Machine and Its Applicability to Recognition Problems", May 2005.
13. Olivier Rochel, « Une approche événementielle pour la modélisation et la simulation de réseaux de neurones impulsionsnels », 2004.
14. Régis MARTINEZ « Apprentissage dynamique dans les réseaux de neurones temporels », 2006.
15. Rudy Guyonneau, Thèse Doctorat « Codage par latence et STDP : des stratégies temporelles pour expliquer le traitement visuel rapide » TOULOUSE, Mars 2006.
16. Yves BURNOD, thèse de doctorat, Modèles Impulsionsnels de Réseaux de Neurones Biologiques, décembre 2003.
17. Loris Nanni, Dario Maio, "Weighted Sub-Gabor for face Recognition", *Pattern recognition letters*, volume 28, Issue 4, March 2007, Pages 487-492.
18. Belhadj S. « Détection et Reconnaissance des Signatures Manuscrites Hors ligne en utilisant les SVM Multi-classes et les Réseaux de Neurones Impulsionsnels » Université USTO Oran, 2007.
19. A. LEHIRECHE. Evolving in Real Time a Neural Net Controller of Robot-Arm: Track and Evolve INFORMATICA, 2004, Vol. 15, No. 1, 63-76 *Institute of Mathematics and Informatics*, ISSN 0868-4952
20. ADJOUDJ Réda, « Authentification Automatique par Identification & Reconnaissance dans un Système de Haute Sécurité », thèse de doctorat en informatique, Université Djillali Liabes de Sidi Bel-Abbès, 2006.

## Fuzzy Set Membership Function Construction Probabilistic Model for Applications in WWW

Tamaz Gachechiladze, Tatiana Kiseliova, Magda Tsintsadze

Ivane Javakhishvili Tbilisi State University  
Institute of Computer Sciences  
Faculty of Exact and Natural Sciences  
University St.2, 0145 Tbilisi, Georgia

[tamaz\\_gachechiladze@rambler.ru](mailto:tamaz_gachechiladze@rambler.ru), [t.kiseliova@web.de](mailto:t.kiseliova@web.de), [magda@parliament.ge](mailto:magda@parliament.ge)

**Abstract.** Fuzzy logic, as a soft computing technique, plays an important role in many Internet applications, as effectively retrieve information from the web, guide users in search, e-commerce with personalized advices to product recommendations, classification and clustering processes. As each fuzzy set is represented by its membership function, the problem to find the appropriate membership function is a main question in any applications. In this paper we propose a new method to solve the problem of a membership function construction that adequately captures the meanings of linguistic terms employed in a particular application. The method is based on the optimal pair of a fuzzy set and its canonically conjugated one.

### 1. Introduction

Recently one of the important application area of fuzzy logic is the Internet, in particular, search and question answering. Fuzzy logic, as a soft computing technique [Za65, Za75, Zi91], plays an important role in such Internet applications, as effectively retrieve information from the web, guide users in search, e-commerce with personalized advices to product recommendations, classification and clustering processes – to mention only a few [SS02].

Fuzzy sets are appropriated to model imprecise and qualitative knowledge, to handle uncertainty. Based on fuzzy logic, the mechanism of approximate reasoning can help in many information retrieval tasks [DM07]. For example, in e-commerce (like automated help in financial process, personalised product recommendation, self-guided shopping process) a translation of user's qualitative preferences of a product to linguistic variables (with their typically multiple linguistic terms) is ideal for best match and personalisation of data [Dy07, Ts07, NW04, GZ07]. In particular, "price" as a linguistic variable can be associated with its linguistic terms such as little, minimal, moderate, considerable, high, very high, extreme. Another example to illustrate fuzzy sets applications in the Internet is the problem to find in the Internet the food with high concentration of vitamin A. Fuzzy sets are inherently suitable for coping with linguistic domain knowledge and producing more interpretable internet solutions [OR06].

Fuzzy logic uses fuzzy rules to track the users and makes the decision about their interests [HHMHL04]. For example, a fuzzy inference model can be employed to establish the relationships between customer needs and alternatives for the multi-functional product: by means of fuzzy rules a customer needs can be translated to the ideal combinations of alternatives. A fuzzy membership function in this case is a way to assign weights to customer needs.

Fuzzy logic is used in web mining, in classification and recommendation and can help to find similar documents, because only for a search with exact words search engines fail to return the desired results [Ts07]. By using the predefined fuzzy clusters, similarity of recommended documents can be presented to a user. Page ranking systems is another possibility of fuzzy logic to show its usefulness in the Internet applications. To deal with overwhelming information when browsing through WWW the personalization that is characteristic feature of fuzzy methodology can be an art of remedy for such diversity [NP05, LDBB04]. Emails filtering and remotely accessing and operating the medical applications are another keywords for the problems in the Internet where uncertainty can arise at the different levels and which can be solved using fuzzy algorithms [LDBB04, TBLA04].

A crucial aspect of each fuzzy logic application no matter where it used is a construction of a membership function. As a fuzzy set is represented by its membership function the problem to find the appropriate membership function is a main question in any applications [KY95].

The question is important because the personalization and formalization of human representation is concentrated mainly in the membership function.

In this paper we propose a new method to solve the problem of constructing membership functions that adequately capture the meanings of linguistic terms employed in a particular application.

The result of intellectual human activity i.e., the controlled inter-action of a subject with an object is almost always incomplete. It is based on limited (generally small) number of attributes (colours) of the object which correspond to the interests of the subject and which he/she can recognize. Sometimes these colours are not available for the direct observation, but are available only in terms of their abstract models (or quantitative models), being the results of the direct perception or some specific measuring procedures. In this case the information loses the definition, univocacy and uncertainty appears there.

Sometimes these abstract colour modes, identifiable on the set of objects, are called variables. When the set of variables are defined as a result of our inter-action with the interesting object, we say that there is defined the system on the object with the given structure of uncertainty. Term “system” is observed as abstract one or as a model of a set of some colours of the object not as a real thing. In other words, the system is a way to look at the object.

Today two sciences study uncertainty systematically:

1. Fundamental sciences of physics are the leading in study of the material (physical) world. There are two types of uncertainties connected with this science: probabilistic (objective) connected with deficiency in empiric information, received by the observation and virtual (intrinsic), quantum, placed in object directly by nature or is the result of the deficiency in means of description language.
2. Fundamental science of informatics is the leading in study of the non-material (informational) world, one of the demonstration of which is the uncertainty connected with the ambiguity and fuzziness. Uncertainty is intrinsic to the expert estimation and to natural language, as the means of description of an informational model.

Data in informatics is a set of so-called informational units. Each of informational units is a four-tuple: (object, sign, value, plausibility) [DP88]. It's important to differ the notion of inaccuracy and uncertainty. Inaccuracy belongs to information content (corresponding to the component “value”), and uncertainty – to its, understandable in terms of compatibility with reality (component “plausibility”). For the given various information there exists the opposition between inaccuracy of expression content and its uncertainty, expressed in that with the increase of expression accuracy, its uncertainty rises as well and vice versa, uncertain character of information leads to some inaccuracy of the final conclusions, received from this information. We see that from one side these notions in a certain way are rival, and from another side – complete each other upon the data presence.

We offer to model this situation by a new concept of optimal pair of a fuzzy set and its canonically conjugated one. Generally, a fuzzy set is constructed on the basis of expert estimations of one of the commutate components. From this point of view, a fuzzy set, constructed in this way, characterizes informational unit incompletely. We offer a method of construction of the membership function of the informational unit taking into account the both conjugate components simultaneously and hence describing this unit in the most complete and optimal way.

## **2. The probabilistic model to build the membership function**

Now we shall shortly consider those general reasons which are laying in the basis of our model construction.

It is significant that every “colour” [Ga96] is of the random quantity and is characterized by a numerical parameter or parameters. Let denote its distribution density by  $\rho(x)$ , it is a nonnegative number and we may present it as  $|\psi(x)|^2$ . (Note that  $\psi(x)$  is an element of Hilbert space  $L^2$ ).

In this case we associate the object with linear operator  $\hat{\wp}$  by the following correspondence:

$$\hat{\wp}(\psi) = x\psi(x),$$

where  $x$  is the value of "colour"'s numerical characteristic, so we may call  $\hat{\phi}$  the "colour operator". Consider Fourier transformation of function  $\psi$  :

$$\phi(y) = \int_{-\infty}^{+\infty} e^{icxy} \psi(x) dx$$

It's known that if  $\int_{-\infty}^{+\infty} |\psi(y)|^2 dx = 1$ , than  $\int_{-\infty}^{+\infty} |\phi(y)|^2 dy = 1$ , so we may consider  $|\phi(y)|^2$  as a density function  $\rho(y)$  for some  $y$  of random quantity.

Thus, every numerical value of "colour" would be in correspondence with random parameter's mathematical expectation value [Ts04].

Analogically with physical colour, which is characterized by the frequency scale, and each value of which is placed in the defined interval, we characterize the generalized colour with a certain interval on  $x$  axis. The analogous picture we have on  $y$  axis (see Figure 1)

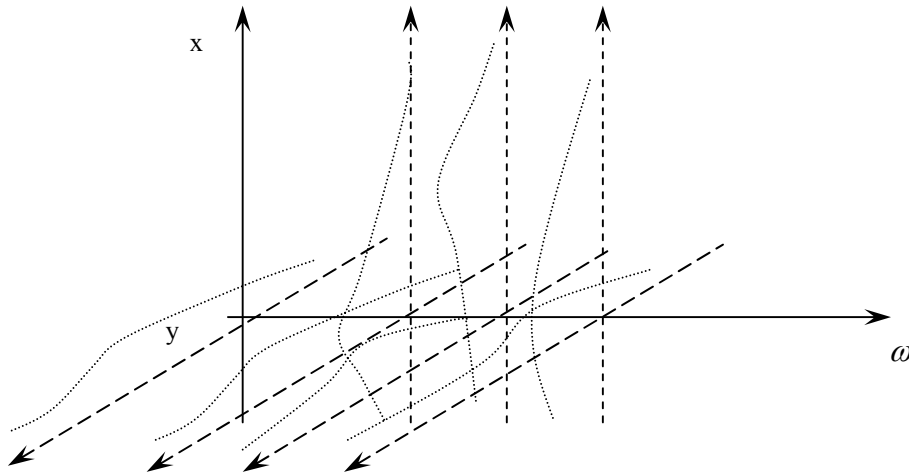


Figure 1.

As  $1 \geq \int_{I_p} \rho(x) dx \geq 0$  we consider it as the value of the membership analogically  $\int_{I_{p_c}} \chi(y) dy$  is the value of other membership function.

We describe the uncertainties on  $x$  and  $y$  axis with appropriate dispersions: it's known that between  $\sigma_x^2$  and  $\sigma_y^2$  exists the following connection:

$$\sigma_x^2 \sigma_y^2 \geq c$$

Thus, we see that, reduction of  $x$  uncertainty value induces increasing in  $y$  uncertainty and vice versa, so if we want to take into consideration both uncertainties simultaneously, we offer the following way.

Let find such  $\rho(x)$  (properly such  $\psi(x)$ ) and  $\chi(y)$  (properly such  $\phi(y)$ ) that  $\sigma_x^2 \sigma_y^2 = c$

Using the analogical method of the quantum-mechanical model solution [Me78], we have:

$$\rho_{\varphi}(x, x^*) = \frac{1}{\sqrt{2\pi\sigma_{\varphi}^2}} \exp\left[-\frac{(x - x^*)^2}{2\sigma_{\varphi}^2}\right] \text{ and } \chi_{\varphi_c}(y, y^*) = \frac{1}{\sqrt{2\pi\sigma_{\varphi_c}^2}} \exp\left[-\frac{(y - y^*)^2}{2\sigma_{\varphi_c}^2}\right]$$

The characteristic intervals will be selected in the following way:

$$I_{\varphi}(\omega) = [x_{\omega}^* - \alpha\sigma_{\varphi}^2; x_{\omega}^* + \alpha\sigma_{\varphi}^2]$$

and

$$I_{\varphi_i}(\omega) = [x_{\omega}^* - \alpha_c\sigma_{\varphi}^2; x_{i\omega}^* + \alpha_c\sigma_{i\varphi}^2]$$

$\alpha$  and  $\alpha_c$  in general cases depend on  $x^*$  and  $y^*$  appropriately. In dispersion terms uncertainty is optimal if

$$\sigma_{\varphi} \cdot \sigma_{\varphi_c} = \text{const}$$

in accordance of this, the membership function is

$$\mu_{\varphi}(\omega) = \frac{1}{\sqrt{2\pi\sigma_{\varphi}^2}} \int_{x_{\omega}^* - \alpha\sigma_{\varphi}^2}^{x_{\omega}^* + \alpha\sigma_{\varphi}^2} \exp\left[-\frac{(x - x_{\omega}^*)^2}{2\sigma_{\varphi}^2}\right] dx$$

$$\mu_{\varphi_c}(\omega) = \frac{1}{\sqrt{2\pi\sigma_{\varphi_c}^2}} \int_{x_{\omega_c}^* - \alpha\sigma_{\varphi_c}^2}^{x_{\omega_c}^* + \alpha\sigma_{\varphi_c}^2} \exp\left[-\frac{(x - x_{\omega_c}^*)^2}{2\sigma_{\varphi_c}^2}\right] dx$$

Thus, we have the membership functions, that correspond to the uncertainty and plausibility, which are optimal (i.e.  $\sigma_{\varphi} \cdot \sigma_{\varphi_c} = \text{const}$ ) according to the above-mentioned scheme. There is a possibility [Me78] to describe both uncertainties simultaneously. The algorithm is as follows: We calculate the "colour pair"  $(\hat{p}, \hat{p}_c)$ 's appropriate numerical parameter (random quantity) membership functions density by the following formula:

$$f(x, y) = \int M(\alpha, \beta) e^{i(\alpha x + \beta y)} d\alpha d\beta$$

Where in the integrated functions  $x$  and  $y$  represent proper values of  $\hat{\varphi}$  and  $\hat{\varphi}_c$  operators, correspondingly, and

$$M(\alpha, \beta) = (\psi^*(x), e^{i(\alpha\hat{p} + \beta\hat{p}_c)} \psi(x)) = (\varphi^*(y), e^{i(\alpha\hat{p} + \beta\hat{p}_c)} \varphi(y))$$

The results of calculations are:

$$f(x, y) = \frac{1}{2\pi} \int_{\mathbb{R}} \psi_{\omega}^* \left( x - \frac{c\tau}{2} \right) e^{-i\tau y} \psi_{\omega} \left( x + \frac{c\tau}{2} \right) d\tau$$

and the membership function is the following:

$$\mu_{\hat{\varphi} \times \hat{\varphi}_c}(x^*, y^*) = \iint_{I(x^*) \times I_c(y^*)} f(x, y) dx dy$$

### 3. Conclusion

In many cases of intellectual human activity, there exists virtually unlimited number of ways of interaction of a subject with an object. This situation is inherent to any Internet application. As each human interaction, this interaction is characterised by uncertainty. To capture this uncertainty we propose a new method to build a membership function by which a fuzzy set is defined. Fuzzy set theory is a known technique of a soft computing that deals with uncertainty.

Our method is based on the concept of an optimal pair of a fuzzy set and its canonically conjugated one. It allows to model inaccuracy and uncertainty that are commutative and completed simultaneous characteristics of an information unit.

Model of canonically complementary (canonically conjugated) subsets can be used in such method of decision-making as discrimination analysis, method of fuzzy probabilities, method of experts, fuzzy differential equations. New approach to the analysis of fuzzy information allows us to control uncertainty, organically inherent to informational units.

### References

- [DM07] M. Detyniecki and C. Marsala. Forest of Fuzzy Decision Trees and Their Application in Video Mining. In Proceedings of the 5th EUSFLAT Conference, Ostrava, Czech Republic, volume 1, pages 345-353, September 11-14, 2007.
- [DP88] D. Dubois and H. Prade. Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York, 1988.
- [Dy07] K. Dyczkowski. A Less Cumulative Algorithm of Mining Linguistic Browsing Patterns in the World Wide Web. In Proceedings of the 5th EUSFLAT Conference, Ostrava, Czech Republic, volume 2, pages 129-137, September 11-14, 2007.
- [Ga96] T. Gachechiladze – Modified Probabilistic Model of Membership Functions of Fuzzy Subsets. Proceedings of Georgian Institute of Applied Mathematics 1(1):63-69, 1996, Tbilisi.
- [GZ07] L. Gu and Y. Q. Zhang. Web Shopping Expert Using New Interval Type-2 Fuzzy Reasoning, Soft Computing, 11:741-751, 2007.
- [HHMHL04] E. Herrera-Viedma, F. Herrera, L. Martinez, J.C. Herrera, and A.G. Lopez. Incorporating filtering techniques in a fuzzy linguistic multi-agent model for information gathering on the web,” Fuzzy Sets and Systems, 143(1):61-83, 2004.
- [KY95] G. J. Klir and B. Yuan. Fuzzy Sets and Fuzzy Logic. Theory and Applications. Prentice Hall PTR, 1995.
- [LDBB04] D. E. Losada, F. Díaz-Hermida, A. Bugarín, and S. Barro. Information access and retrieval (IAR): Experiments on using fuzzy quantified sentences in adhoc retrieval. In Proceedings of the 2004 ACM symposium on Applied computing SAC '04, ACM Press, March 2004.
- [Me78] A. Messiah. Quantum Mechanics (in Russian). Moscow, Nauka, 1978.
- [NP05] B. L. Narayan and Sankar K. Pal. Fuzzy Web Server Model. In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005, pages 1-4.
- [NW04] E. W. T. Ngai and F. K. T. Wat. Fuzzy decision support system for risk analysis in e-commerce Development. Decision Support Systems, 40: 235-255, March 2004.
- [TBLa04] A. Taleb-Ahmed, A. Bigand, V. Lethuc, and P.M. Allieux. Visual acuity of vision tested by fuzzy logic: An application in ophthalmology as a step towards a telemedicine project. Information Fusion, 5(3):217-230, September 2004.]
- [Ts07] C. Tseng. Internet Applications of Fuzzy Logic and Neural Networks : A Survey. Engineering, Soft Computing and Applications, vol. 1, issue 2, 2007.
- [OR06] A.C. Ojha, S. K. Radhan. Fuzzy Linguistic Approach to Matchmaking in Ecommerce. In Proceedings of 9th International Conference on Information Technology (ICIT'06), December 2006, pages 217-220.
- [SS02] E. Sanchez and J.-A. Santini. Fuzzy Logic and the Internet: a Fuzzy Web Site Evaluator, Annual Meeting of the North American Fuzzy Information Processing Society, 2002, pages 344 – 347.
- [Ts04] M. Tsintsadze – Probabilistic Model of Optimal Membership Function. Proceedings of Iv. Javakhishvili Tbilisi State University, 354(24): 75-84, 2004, Tbilisi
- [Za65] L. A. Zadeh. Fuzzy sets. Inf. Control, 8:338-353, 1965.
- [Za75] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning, Part I, Part II, Part III. Information Sciences, 8,9:199-249, 301-357, 43-80, 1975.
- [Zi91] H.-J. Zimmermann. Fuzzy Set Theory and Its Applications. Kluwer Academic Publishers, 3300 AH Dordrecht, The Netherlands, 1991



*Posters*

---



## **Towards new generation of information search tools on the web using ontologies**

A.Raouf Kouhoul

Department of Computer Science, A.Mira University, Bejaia, 06000 Algeria  
Kouhoul\_abderaouf@yahoo.fr

**Abstract.** Searching the needed piece of information becomes more and more a hard task facing the immense growth of the volume of information available on the web in different forms: plain text, sound, video. By using information search tools (search engine, directories, intelligent agents and meta-search engines...) the results are in the most cases less accurate; this lack is due to the classic web structure which is designed to be human rather than machine comprehensible. Facing this situation, a need of accuracy information search tools has appeared. The main idea of our work is to use ontologies in the reformulation of user queries submitted to search tools by the context of the concepts included within the query. For this purpose, we have developed a meta-search engine that uses this technic in order to get more accurate results.

**Keywords:** Information search tools, meta-search engine, Ontologies, semantic web.

## **Proposition d'un modèle pour Ordonnancement d'un Système Automatisé de Production par Intensification de Chromosomes**

Djamila Bouhalouan<sup>1</sup>, Nassima Aissani<sup>1</sup>, Bouziane Beldjilali<sup>2</sup>

<sup>1</sup> Département d'informatique, Faculté de Sciences, Université d'Oran Es-Senia,  
BP 1524, El-M'Naouer, 31000, Oran, Algérie  
{jjilibin,aissani.nassima,bouzianebeldjilali}@yahoo.fr

**Résumé.** Le principe d'un ordonnancement coopératif selon une modélisation multi-agents résulte du rapprochement des domaines de l'Ordonnancement et de l'Intelligence Artificielle Distribuée (IAD). Nous avons visé l'une des métaheuristiques avancées Algorithmes Génétiques Hybrides (AGH) dans notre SMA, après avoir effectué une comparaison entre AG Simples et AG Hybrides, et d'étudier la contribution de ces métaheuristiques dans l'amélioration des performances du système de production, et pour l'évolution génétique des agents, des groupes et par extension des organisations. Le présent article, explique notre démarche pour réaliser un système d'ordonnancement d'un atelier Flow-Shop en se basant sur une approche multi-agents et plus spécifiquement des agents génétiques. La démarche de résolution est une hybridation entre les opérateurs génétiques et une recherche locale pour un but d'exploration efficace de l'espace de recherche (Intensification), et pour recentrer la communauté des agents autour d'une solution optimale. Une réalisation de ce modèle se présentera en fin de l'article.

**Mots Clés :** Ordonnancement, Système de Production, Multi-Agents, Algorithmes Génétiques, Recherche Locale.

## **Modélisation d'un système d'information assurant l'interopérabilité des appareils mobile**

Makhlouf Derdour<sup>1</sup>, Nacira Ghoualmi-Zine<sup>2</sup>, Philippe Roose<sup>3</sup>

<sup>1</sup> Université de Tébessa – Algérie, Département d'informatique - Tébessa – Algérie  
[m.derdour@yahoo.fr](mailto:m.derdour@yahoo.fr)

<sup>2</sup> Université Badji Mokhtar – Algérie, Département d'informatique - Annaba – Algérie  
[ghoualmi@yahoo.fr](mailto:ghoualmi@yahoo.fr)

<sup>3</sup> Université de Pau, Château-Neuf - Place Paul Bert - Bayonne - FRANCE  
[Philippe.Roose@iutbayonne.univ-pau.fr](mailto:Philippe.Roose@iutbayonne.univ-pau.fr)

**Résumer.** L'objectif de cet article est de proposer un système d'information capable de régler les problèmes d'adaptations au niveau des appareils mobiles, ce système est constitué d'un ensemble de processus, chaque processus assure une fonction bien précise. Pour chaque type d'adaptation on fait appel à plusieurs processus, le choix des processus dépend de la qualité de service qu'on veut avoir (la vitesse d'adaptation, le taux de compression, la qualité des composants d'un message,...etc).

**Mots clés :** Multimedia Messaging Service (MMS), Média, Adaptation, Transformation, Appareil mobile, hétérogénéité.

## **Le Multicast dans un Réseau Cellulaire**

Nassima Adjissi<sup>1</sup>

<sup>1</sup> Département Informatique, UFA, Sétif, Algérie  
[nadjissi@hotmail.com](mailto:nadjissi@hotmail.com)

**Résumé.** Dans ce document nous présentons une étude synthétique des différents travaux de recherche s'intéressant à la fiabilité de communication de groupe (Multicast ou diffusion sélective) dans un réseau cellulaire. Du fait des modifications rapides de la composition des cellules, et aux temps de latence dans le réseau fixe, il est possible pour un mobile de rater un message diffusé, étant sorti d'une cellule avant que celle-ci ait reçu le message et entré dans une cellule après que celle-ci ait reçu ce même message. De même, il est possible qu'un mobile reçoive plusieurs fois un message identique. Ainsi après avoir analysé les caractères spécifiques des réseaux cellulaires et l'avantage et limitation de différentes techniques et stratégies adoptées par les protocoles courants de Multicast, nous proposons un protocole fiable de Multicast en améliorant un des protocoles existant.

**Mots clés:** Réseaux Cellulaires, Calcul Mobile, Communication Multidestinataire, Groupe Multicast, Handoff.

## **Informatique Mobile : localisation des utilisateurs**

Zohra SLAMA

Département d'informatique  
Faculté des Sciences de l'Ingénieur  
Université Djillali Liabés  
Sidi-Bel-Abbés (22000) ALGERIE  
zslama@yahoo.com

**RÉSUMÉ.** Dans un réseau mobile, les utilisateurs qui disposent d'unités mobiles peuvent se déplacer d'une localisation à une autre quittant ainsi leur réseau d'attache. Malgré cela, il est nécessaire de donner la possibilité aux utilisateurs d'exploiter les ressources du réseau durant leurs déplacements et ce sans contraintes géographiques. Pour atteindre un tel objectif, la localisation des utilisateurs est une information pertinente. Dans le cas des réseaux mobiles, cette localisation ne peut pas être obtenue à partir de l'identificateur du mobile comme dans un réseau fixe, où l'information de routage se trouve dans l'adresse de chaque nœud du réseau. Ainsi, la mise en place d'un service de gestion de la localisation des mobiles devient indispensable. Dans ce papier, nous proposons une approche parallèle pour localiser des utilisateurs dans un réseau mobile. Cette approche est basée sur une stratégie de parcours en largeur de l'arbre représentant un réseau.

**MOTS-CLÉS :** Réseau mobile, Cellule, Localisation, Base de données, Réseau hiérarchique, Optimisation, Requêtes, Parallélisme.

## **Electronic corpora and computational linguistics**

Mohamed Grazib\*, Bel Abbés Neddar \*\*

\*Djillali liabes university : Computer science institute .  
Sidi Bel Abbas. Algeria

\*\*University of Mostaganem. Algeria

\*mfgrazib@hotmail.com

**Abstract:** the uses of electronic corpora in computational linguistics are of great importance. Nowadays it is possible to study linguistic phenomena by using statistical analyses. Concordances, collocations and frequencies are the key elements for making linguistic researches more available, more adequate and more accurate.

**Keywords:** corpus, computational linguistics, corpus linguistics.

## **Un Méta-Modèle Ontologique pour la Représentation de Domaines d'Enseignement en E-learning**

Farida Bouarab-Dahmani\* — Malik Si-Mohammed\*\*

Département Informatique,  
Université Mouloud Mammeri de Tizi-Ouzou (15000) Algérie  
\*dahmani\_farida@yahoo.fr, \*\*maliksim@hotmail.com

**Résumé.** Ce papier présente un méta-modèle pour la représentation de domaines d'enseignement sous forme d'une ontologie générique avec une proposition d'opérationnalisation partielle basée sur le langage XML. La problématique traitée se rapporte à la représentation de domaines d'enseignement en e-learning en tenant compte de leurs structures, leurs didactiques d'une part et de leurs présentations et formats d'échanges via le Web d'autre part. L'introduction d'un méta-modèle est dans l'objectif de faciliter la représentation et la manipulation de plusieurs domaines d'enseignement avec le même langage. L'ontologie générique a fait objet de deux exemples d'instanciations relativement aux domaines d'enseignement de l'algorithmique et des bases de données relationnelles.

**Mots Clés :** Modélisation de domaine d'enseignement, ontologies, XML, enseignement de l'algorithmique, enseignement des bases de données relationnelle.



## **Opération de télé-enseignement en Algérie: Vers quelle pratique pédagogique? Classique ou socioconstructiviste?**

Hadda CHERROUN<sup>1</sup>, Mouloud BOUZOUAD<sup>1</sup>

<sup>1</sup> Université Amar Telidji Laghouat,  
{hadda\_cherroun, m.bouzouad}@mail.lagh-univ.dz

**Résumé:** Le ministère algérien de l'enseignement supérieur et de la recherche scientifique, dans le cadre d'un programme de soutien au développement des TICE, a initié un projet de téléenseignement. Dans cet article, nous essayons d'analyser en particulier l'opération de visioconférence de ce projet en tant qu'innovation pédagogique. Le but essentiel de cette étude est d'identifier les types d'usage et les pratiques de cette potentialité, à mettre en place et qui sont susceptibles de favoriser un modèle d'apprentissage et de travail collaboratifs. Notre méthodologie de caractérisation se base sur l'observation des étapes de mise en place et d'installation du dispositif de visioconférence ainsi que sur l'analyse critique des différents choix faits (caractéristiques techniques du matériel, répartition des sites émetteur /récepteur de visioconférence ainsi que les mesures d'accompagnements prévues pour la réussite de cette opération) Nous terminons par proposer un ensemble de mesures d'accompagnement à entreprendre permettant de piloter, d'orienter la genèse d'un cadre d'usage, de cet artefact, qui peut être adapté au travail collaboratif et de capitaliser l'investissement réalisé pour l'acquisition de ce dispositif, déjà là !

**Mots clés:** Téléenseignement, Visioconférence, Travail collaboratif, modèle Socioconstructiviste, Usages.

## **An Agent-based Architecture for an Intelligent Decision Support System: Application to distributed real-time Scheduling**

Noria Taghezout<sup>1</sup>, Karim Bouamrane<sup>1</sup>

<sup>1</sup>Computer Science Department, University of Es-Senia Oran, BP 1524, El-M' Naouer,  
31000, Oran, Algeria  
taghezoutnour@yahoo.fr, kbouamranedz@yahoo.fr

**Abstract.** Two factors have changed the Decision support nature and still dominate the IDSSs (Intelligent Decision support System) future: (1) global enterprises that needs distributed information and fast decision making; and (2) the internet which enables access to distributed information and speed. Agents have permitted response to these IDSSs requirements. Intelligent software agents do not only provide an approach for functional integration, but also define collaboration among enterprises by using autonomous intelligent decision making. This paper proposes a generic model for distributed real-time scheduling based on multi agent architecture. Agents provide some actions or services including data collected from heterogeneous sources, information integration and analysis, real time scheduling and decision making. The decision centers negotiate compromises for conflict solving. The framework embeds expert knowledge within the DSS. It will rely on the use of ontologies to structure data access. The approach is tested for simple scenarios.

**Keywords:** Multi agent System (MAS), Negotiation, Intelligent Decision Support System (IDSS), Distributed real- time Scheduling, Production System (PS).

## **Learning Domain Ontology for Semantic Web Services**

Djamila Berramdane<sup>1</sup>, Djamel Amar Bensaber<sup>2</sup>, Mimoun Malki<sup>1</sup>

<sup>1</sup> EEDIS Laboratory, Computer Science Department, University of Sidi Bel Abbes,  
Algeria

<sup>2</sup> Liris Laboratory, Claude Bernard University, 8 Boulevard Niels Bohr, 69622 Villeurbanne,  
cedex France

{Djamila Berramdane, [berramdane\\_mail@yahoo.fr](mailto:berramdane_mail@yahoo.fr)}, {Djamel Amar Bensaber,  
[djamel.amar-bensaber@liris.cnrs.fr](mailto:djamel.amar-bensaber@liris.cnrs.fr)}, {Mimoun Malki, [m\\_malki@yahoo.com](mailto:m_malki@yahoo.com)}

**Abstract.** In order to answer the current problems of the Web services, several recent works proposes the use of Domain ontologies to automate the task of invocation and composition. However seen difficulty of construction of these domain ontologies it became necessary to adopt the approaches developed in the fields of ontology learning and adjusted them in the context of the Web services, this context is characterized by the strong use of the language XML to specify the data exchanged by the service Web. In this article we propose an approach of learning domain ontology of Web service from the xsd ComplexType and xsd SimpleType specified in WSDL file. Using a set of transformation rules of the specification hierarchy of the data defines by the xsd schema a semantic schema is generated, after that we enriched this schema by new relations semantics extracted using a set of Patterns.

**Keywords:** Ontology Learning, semantic web services, learning domain ontologies for web services.

## **Optimisation de la consommation d'énergie avec control de topologie dans les réseaux de capteurs sans fil**

K.FELLAH, B.KECHAR

Université d'Oran Es-Sénia, faculté des sciences, département d'Informatique  
BP 1524 EL M'NAOUAR, ORAN, ALGERIE  
{fellaikhadjia, bkechar2000}@yahoo.fr

**Résumé:** Dans cet article nous présentons un programme linéaire (PL) d'optimisation en variables 0-1 dont le but est d'économiser de l'énergie et par conséquent prolonger la durée de vie d'un Réseau de Capteurs Sans Fil (RCSF), en faisant déplacer certains noeuds capteurs. L'idée principale sur laquelle repose notre démarche d'optimisation consiste à déplacer quelques noeuds capteurs avant les transmissions de données au lieu de réaliser ces transmissions directement après déploiement. Ces déplacements sont effectués en préservant des contraintes liées à la topologie du RCSF : couverture et connectivité. Les résultats préliminaires obtenus par expérimentation sous AMPL montrent que cette technique d'optimisation permet à la fois d'économiser de l'énergie globale du réseau et de préserver le contrôle de topologie de celui-ci.

**Mots clé:** RCSF, Programmation linéaire en variables 0/1, contrôle de topologie, optimisation de la consommation d'énergie.

## **Semantic Web Services: A reverse engineering approach for the WSMO Specifications**

Houda El Bouhissi<sup>1</sup>, Mimoun Malki<sup>2</sup>, Djelloul Bouchiha<sup>3</sup>

<sup>1</sup> EEDIS Laboratory, University of Sidi Bel Abbes 22000, Algeria  
[Houda.elbouhissi@gmail.com](mailto:Houda.elbouhissi@gmail.com)

<sup>2</sup> EEDIS Laboratory, University of Sidi Bel Abbes 22000, Algeria  
[Malki@univ-sba.dz](mailto:Malki@univ-sba.dz)

<sup>3</sup> EEDIS Laboratory, University of Sidi Bel Abbes 22000, Algeria  
[bouchiha.dj@gmail.com](mailto:bouchiha.dj@gmail.com)

**Abstract.** Semantic web services are at the convergence of two research fields related to Internet technologies, namely semantic Web and Web services. Several solutions were proposed for the specification of the semantic web services such as OWL-S (Ontology Web Language for services) and WSMO (Web Service Modeling Ontology). However, these technologies require a human user intervention to select and choose the required services. We propose an approach to create service ontology with WSMO specifications starting from WSDL (Web Service Language Description), SOAP (Simple Object Access Protocol) and UDDI (Universal Description, Discovery and Integration) by using the reverse engineering techniques. Reverse engineering aims to extract several types of information of existing software and to employ them for comprehension, renovation as well as maintenance of the system.

**Keywords:** Ontology, Semantic Web, Semantic Web Services, Web Services, WSDL, SOAP, UDDI, Reverse engineering.

## **Des Mécanismes de Tolérance aux Fautes pour un Réseau de Capteurs Sans Fil**

Benahmed Khelifa<sup>1</sup>, Prof. Merabti Madjid<sup>2</sup>, Prof. H. Haffaf<sup>3</sup>, Laarbi Abdelmadjid<sup>4</sup>

<sup>1</sup> Institut des sciences exactes, Centre universitaire de Béchar, benahmed\_khelifa@yahoo.fr

<sup>2</sup> School of Computing and Mathematical, LJMU, U.K

<sup>3</sup> Faculté des sciences, Département d'Informatique, Université d'Oran Es-Sénia, haffaf\_hafid@yahoo.fr

<sup>4</sup> Institut des sciences exactes, Centre universitaire de Béchar, amdlarbi@yahoo.fr

**Résumé:** Il existe plusieurs types de réseau sans fil: réseau ad hoc, réseau maillé, réseau de capteurs, etc... Mais tous cherchent à s'auto organiser et s'auto configurer que ce soit pour améliorer les performances, économiser de l'énergie et des ressources ou encore pour acheminer des données. Cependant, les entités de ces réseaux sont mobiles, sans infrastructure fixe, où la panne et très fréquentes (problèmes d'énergie, vulnérabilité aux attaques, et de topologie dynamique, etc ...), et peuvent également apparaître et disparaître de façon arbitraire, tout ceci provoquant des changements de la topologie du réseau. L'objectif de cet article sera la proposition de mécanismes tolérants aux fautes et aux malveillances dans un réseau de capteurs sans fil. L'étude comportera la proposition de mécanismes originaux pour la tolérance aux fautes dans ce type de réseau. Toutes ces études pourront être menées aussi bien par simulation, expérimentation que par analyse théorique.

**Mots clés :** systèmes mobiles, réseaux de capteurs sans fil, tolérance aux fautes, auto organisation.